frontiers | Research Topics

Current thoughts on the brain-computer analogy - all metaphors are wrong, but some are useful

Edited by
Giorgio Matassi, Pedro Martinez and Bud (Bhubaneswar) Mishra

Published in
Frontiers in Ecology and Evolution
Frontiers in Computer Science

ISSN 1664-8714
ISBN 978-2-83251-651-5
DOI 10.3389/978-2-83251-651-5

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Current thoughts on the brain-computer analogy - all metaphors are wrong, but some are useful

Topic editors

Giorgio Matassi — FRE3498 Ecologie et dynamique des systèmes anthropisés (EDYSAN), France

Pedro Martinez — University of Barcelona, Spain

Bud (Bhubaneswar) Mishra — New York University, United States

Citation

Matassi, G., Martinez, P., Mishra, B., eds. (2023). *Current thoughts on the brain-computer analogy - all metaphors are wrong, but some are useful*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83251-651-5

Table of contents

Frontiers in Ecology and Evolution

TYPE Editorial
PUBLISHED 01 February 2023
DOI 10.3389/fevo.2023.1130510

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Gordana Dodig-Crnkovic,
Chalmers University of Technology, Sweden

*CORRESPONDENCE
Giorgio Matassi
✉ giorgio.matassi@uniud.it
Pedro Martinez
✉ pedro.martinez@ub.edu
Bud Mishra
✉ mishra@nyu.edu

SPECIALTY SECTION
This article was submitted to
Models in Ecology and Evolution,
a section of the journal
Frontiers in Ecology and Evolution

RECEIVED 23 December 2022
ACCEPTED 11 January 2023
PUBLISHED 01 February 2023

CITATION
Matassi G, Mishra B and Martinez P (2023)
Editorial: Current thoughts on the brain-computer analogy—All metaphors are wrong, but some are useful.
Front. Ecol. Evol. 11:1130510.
doi: 10.3389/fevo.2023.1130510

Editorial: Current thoughts on the brain-computer analogy—All metaphors are wrong, but some are useful

Giorgio Matassi[1,2]*, Bud Mishra[3,4]* and Pedro Martinez[5,6]*

[1]Université de Picardie Jules Verne, UMR "Ecologie et Dynamique des Systèmes Anthropisés" (EDYSAN, UMR 7058 CNRS), Amiens, France, [2]Dipartimento di Scienze AgroAlimentari, Ambientali e Animali, University of Udine, Udine, Italy, [3]Courant Institute of Mathematical Sciences, New York University, New York, NY, United States, [4]Cold Spring Harbor Laboratory, New York, NY, United States, [5]Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Barcelona, Spain, [6]Institut Català de Recerca I Estudis Avançats (ICREA), Barcelona, Spain

KEYWORDS

philosophy of science, semantics, metaphor, analogy, Computer Science, information, Neuroscience, Neural Networks

Editorial on the Research Topic
Current thoughts on the brain-computer analogy—All metaphors are wrong, but some are useful

This project kicked off in the fall of 2020. There are two parts of the title of this Research Topic—Special Issue. The first one evokes the issue raised by Turing ("*Can machines think?*", Turing, 1950), a question that we, the Editors, revisit reflecting our complementary multi-disciplinary backgrounds (Evolutionary Biology, GM; Evo-Devo, PM; and Computer Science, BM) and take it up again with a fresh start; this question made us realize how ripe the Brain-Computer analogy has become for a reassessment. The complexity of the subject needed the involvement of experts from the different fields that have been concerned with many related problems, namely Natural Sciences (here Biology and Physics), Mathematics, Psychology and Philosophy. Indeed, the Topic is certainly timely for, while this Issue was going to press, a number of publications have appeared that tackle these very issues both in Sciences (Reynolds, 2022; Yang and Lu, 2022) and Humanities (Kelty-Stephen et al., 2022).

The second part of the title paraphrases a well-known aphorism in Statistics: "*Essentially, all models are wrong, but some are useful*" (George E. P. Box). This statement introduces the "philosophical" part of this topic, *viz.* the semantic issue; in Turing's words: "*Can machines think? This should begin with definitions of the meaning of the terms 'machine' and 'think'*" (Turing, 1950). Indeed, both the Authors and the Editors of this Special Issue realized that a number of other concepts, crucial to evaluate the Brain-Computer relationships, were in need of an updated definition: machine (Bongard and Levin), computer (Danchin and Fenton; Richards and Lillicrap), metaphor and analogy (Matassi and Martinez).

We started off by making a wishful list of relevant topics that would embrace a vast a spectrum of disciplines concerned. These were: Brain architecture, evolution and functioning; Neural Networks and Computational Neuroscience; *Network Science* (*network evolution*); Computer Science; Information theory; Artificial Intelligence (AI); *Game theory; Quantum* *brain—quantum computer*; *Evo-Devo*; *Neurobiology Experimental research*. In so doing, we hoped to stimulate a multi-, trans-, and inter-disciplinary authorship of the articles in this Research Topic, though we were fully aware of the controversies and debates that could arise among scientists and technologists from such diverse scientific backgrounds. Unsurprisingly, we did not fully succeed, and a number of relevant topics had to be left missing from this SI (in *Italics* in the list above).

Many other equally important disciplines were not included for they would each deserve a full Research Topic: Consciousness/Mind, Cognition, Behavior, Language, and Culture.

We have subjectively subdivided the articles in the Special Issue into five subject-wise sections, following some close relationships in their contents. Our groupings are (more details below): 1-Historical Perspectives (Cobb), 2-Philosophical Implications (Brette; Chirimuuta; Gomez-Marin), *3*-Utility and limitations of the brain-computer metaphor (Bongard and Levin; Danchin and Fenton; Davis; Fraser et al.; Richards and Lillicrap; Roli et al.), 4- Extending the concept of cognition (Gershenson), and 5-A new metaphor for the brain, the internet (Graham).

The problem area is introduced in a first paper co-authored by Matassi and Martinez (two of the three editors of this SI). The authors introduce the Research Topic and provide a detailed review of the other 12 contributions; this is complemented by a graphical summary linking articles to selected concepts. Moreover, they analyze in detail the distinction between metaphor and analogy, and offer a definition for the latter. They introduce the notion of Brain and the related evolutionary theories. The article closes with thoughts on creativity in Science, for … "if we ask "can computers think," next we ought to ask "can computers create." And the very act of creation (be it in sciences or in the arts) stems from the awareness of the aesthetic element."

Before summarizing the papers included in this SI, let us consider, briefly, what is the problem area we are trying to deal with in this issue. This introspection should provide us with a reference mark in which the discussion takes place. Obviously, we need to start by understanding what a metaphor is and what purposes it serves, with the emphasis in one of the most productive metaphors in science, the "Brain as a Computer." History tells us that the metaphor has been enriched or modified over time, incorporating new concepts arising in different disciplines, from neuronal physiology to circuit assemblies, information processing and the genesis of complex systems.

1. Historical perspectives

The revolutionary studies of Cajal and Golgi brought us a completely new view of the brain as a biological tissue. The intricate nature of its unit connections (neurons and substructures) suggested the possibility that the brain is actually a connected set of wires, with complex architectures. Moreover, the discovery of chemical and electrical connections between neurons reinforced the image of a giant electrical device with multiple, complex, switching mechanisms. The emergence of the information age, with the first devices able to "compute" operations, was instrumental in bringing a new model of the brain, understood as a complex computing device able to perform logical functions. The history of some old and new metaphors for the brain are nicely exposed by Cobb. This article introduces, from a historical perspective, the current debates in the field, as reflected in the next series of articles in this SI.

2. Philosophical implications

Metaphors are considered either as linguistic (semantic) or cognitive devices, rooted in concrete brain structures, that help us navigate the world. More than this, they help translate complex descriptions into less cognitively demanding ones. Much research is being conducted into the neurobiological basis of metaphoric thinking, but this is a problem we will not touch on in this introduction (see Gomez-Marin's paper for further commentary). As in other complex systems (e.g., the structure of the universe, the prediction of weather or the behavior of large social groups), the study of the brain has been subject to a series of reductionist descriptions.

In a suggestive paper, Chirimuuta comments on the assertion by different authors that have hypothesized the brain and computers (or any other complex artifact) as tractable using multi-level approaches. However, as appealing the simile can be, Chirimuuta thinks that there are several limitations that need to be accounted for, and she provides us with a careful discussion of all of those. In a similar line, a major concern of Brette's is "*What is a computer*?" This is followed by a reappraisal of the concept of "*program*." In this context he discusses the notions of algorithm and computation in the brain, and from a philosophical perspective he asks: "what is a *brain program*"? and, if true, "*who gets to 'program' the brain?*". All those papers bring us to the fundamental role of introducing concepts in our discussions, to make them meaningful. From the very concept of a metaphor to what actually would do a "computerized" brain, all contribute to clarify the terms of discussion.

3. Utility and limitations of the brain-computer metaphor

Whether a metaphor has a practical utility depends very much on what predictions it makes and how valid are the assumptions that underlie the use of these metaphors. In a series of papers, we are confronted with the idea of how computers (or its derivative AI technologies) can imitate humans, or certain human capacities. While the Bongard and Levin view is certainly optimistic, assuming that modern/future machines can actually imitate humans, Danchin and Fenton; Fraser et al.; Roli et al. point to some irreducible properties that make the human mind, essentially, inimitable, thus stating in different ways that brains are not digital computers. Davis takes a more neutral position and just ask himself whether this is a realizable possibility or not.

4. Extending the concept of cognition

When discussing the human mind, two concepts are normally mentioned, that of "intelligence" and "cognition." In an interesting article Gershenson revisits the concept of intelligence as the result of brain information processing. He suggests to use measures of information as a tool to study cognitive systems, including brains and computers. In addition, suggests looking at cognition beyond the individual, and analyze cognition in collectives such as insects' swarms.

5. A new metaphor for the brain, the internet

More recently, some authors have pointed out the need to incorporate the problematics of information flow and storage in the brain within our models. Others have pointed to the idea that our brains might be using quantum processes in cognition and consciousness (Kerskens and López-Pérez, 2022). This has prompted the incorporation of other, recent, models of the brain such as the internet (Graham, this issue).

6. General conclusion

The use of metaphors has helped both scientists and lay men to understand the brain functions in history. Whether they have been accurate or misleading, they have provided us with tools to interrogate the brain. Every period has had their favorite metaphor to explain how this organ performs functions that allow us to both comprehend and navigate the world. These metaphors have an intricate relationship with current ideas of the matter and the living, in most cases suggesting different mechanical similes for our organs (i.e., the brain). Among the mechanical metaphors, arguably, a very productive one has been that of equating brain with a computer device. In recent decades this very last image has changed (as always did), in parallel with our refinement of computing technologies. The arrival of AI, with the possibilities (sometimes overblown) for imitating humans, has generated a heated debate on whether machines can imitate, and perhaps substitute, humans in all endeavors, including those that rely on creativity. We should expect that in the next few years, our view of the brain as a computer will be completely transformed, perhaps entering into an era in which brain and computer performances will be indistinguishable. We asked ChatGPT (a natural language processing tool driven by AI technology) for an opinion: It agreed imitatively and wrote: "*The brain-computer analogy is a metaphor that has been used extensively in the field of artificial intelligence and cognitive science to describe the relationship between the human brain and computers. The metaphor suggests that the brain can be thought of as a computer, and that the way in which it processes information and solves problems can be understood in terms of computer algorithms and hardware, Etc.*" Clearly, a rather cheap imitation of this Special Issue's esteemed editors!

Author contributions

All authors contributed to the writing of this Editorial.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Kelty-Stephen, D. G., Cisek, P. E., De Bari, B., Dixon, J., Favela, L. H., Hasselman, F., et al. (2022). In search for an alternative to the computer metaphor of the mind and brain. *arXiv:2206.04603* doi: 10.48550/arXiv.2206.04603

Kerskens, C. M., and López-Pérez, D. (2022). Experimental indications of non-classical brain functions. *J. Phys. Commun.* 6, 105001. doi: 10.1088/2399-6528/ac94be

Reynolds, A. S. (2022). *Understanding Metaphors in the Life Sciences (Understanding Life)*. Cambridge: Cambridge University Press.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind* 59, 433–460. doi: 10.1093/mind/LIX.236.433

Yang, H., and Lu, Z. (2022). "Computerizing connections between creativity and aesthetics," in *2022 IEEE International Conference on Service-Oriented System Engineering (SOSE)* (Newark, CA: IEEE), 185–188. doi: 10.1109/SOSE55356.2022.00028

frontiers | Frontiers in Ecology and Evolution

TYPE Review
PUBLISHED 13 January 2023
DOI 10.3389/fevo.2022.1099253

The brain-computer analogy—“A special issue”

Giorgio Matassi[1,2]* and Pedro Martinez[3,4]*

[1]UMR "Ecologie et Dynamique des Systèmes Anthropisés" (EDYSAN, UMR 7058 CNRS), Université de Picardie Jules Verne, Amiens, France, [2]Dipartimento di Scienze AgroAlimentari, Ambientali e Animali, University of Udine, Udine, Italy, [3]Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Barcelona, Spain, [4]Institut Català de Recerca I Estudis Avançats (ICREA), Barcelona, Spain

OPEN ACCESS

EDITED BY
Thilo Gross,
Helmholtz Institute for Functional Marine Biodiversity (HIFMB), Germany

REVIEWED BY
Laura Antonia Barth Martinez,
Helmholtz Institute for Functional Marine Biodiversity (HIFMB), Germany
John Bickle,
Mississippi State University,
United States
Soumyajit Mandal,
Brookhaven National Laboratory (DOE),
United States
Marcos Cramer,
Technical University Dresden,
Germany
Michael Levin,
Tufts University,
United States

*CORRESPONDENCE
Giorgio Matassi
✉ giorgio.matassi@uniud.it
Pedro Martinez
✉ pedro.martinez@ub.edu

SPECIALTY SECTION
This article was submitted to Models in Ecology and Evolution, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 15 November 2022
ACCEPTED 23 December 2022
PUBLISHED 13 January 2023

CITATION
Matassi G and Martinez P (2023) The brain-computer analogy—“A special issue”. *Front. Ecol. Evol.* 10:1099253.
doi: 10.3389/fevo.2022.1099253

In this review essay, we give a detailed synopsis of the twelve contributions which are collected in a Special Issue in Frontiers Ecology and Evolution, based on the research topic “Current Thoughts on the Brain-Computer Analogy—All Metaphors Are Wrong, But Some Are Useful.” The synopsis is complemented by a graphical summary, a matrix which links articles to selected concepts. As first identified by Turing, all authors in this Special Issue recognize semantics as a crucial concern in the brain-computer analogy debate, and consequently address a number of such issues. What is missing, we believe, is the distinction between metaphor and analogy, which we reevaluate, describe in some detail, and offer a definition for the latter. To enrich the debate, we also deem necessary to develop on the evolutionary theories of the brain, of which we provide an overview. This article closes with thoughts on creativity in Science, for we concur with the stance that metaphors and analogies, and their esthetic impact, are essential to the creative process, be it in Sciences as well as in Arts.

KEYWORDS

neuroscience, computer science, semantics, metaphor, analogy, creativity, philosophy of science

1. Introduction

Drawing comparisons between Brains and Computers has been a long intellectual exercise carried out by Philosophers, Psychologists, Mathematicians, Physicists, Computer Scientists and Neuroscientists. While some authors have suggested that this is a vacuous discussion (they assume that brains are “obviously” computers), others believe that there are instances in the functioning of both systems that do not allow this easy jump to conclusions, meriting further analysis. Some of the confusions come from the fact that, according to some authors many researchers do not understand the fact that behaving intelligently does not mean being just an information processor (because computers seem to behave intelligently using processors it does not mean that “intelligence” and “information processing” are equivalent; opinion articulated by psychologist Robert Epstein assay “The empty brain” 2016). Others, however, think that the assignment of a name such as “computational system” to the brain is limiting and biases the way in which we see brain processes occurring, e.g., consciousness, awareness, or simply “making sense of the world.”

In this context, we think that re-visiting the Brain-Computer analogy is still a very valid endeavor. Indeed, based on the Research Topic "Current Thoughts on the Brain-Computer Analogy—All Metaphors Are Wrong, But Some Are Useful" this Special Issue gathers 12 articles that deal with these problematics, all showing that the subject (or the debate) is still very much alive.

1.1. The research topic

From the very beginning, it was clear to us that the project was to be constructed according to the architecture of a network, since the main concepts around which it was conceived were interconnected at various degrees. We reasoned that the structure of such a network would favor *"information"* to be exchanged among nodes (disciplines, approaches, articles). This useful approach allowed us to view our subject (brain) and the approaches taken to study it as two manifestations of a similar phenomenon, and the reticular interconnection of nodes (ideas, approaches or physical entities) as the graphical expression of these connections.

The topics and concepts that we saw as the *"nodes of the network,"* and which formed the foundation of the project, were the following: Conceptual points (Philosophy); Network Science; Complex systems (self-organization); Neural Networks and Computational Neuroscience (*Artificial neural networks*); Computer Science (*distributed-centralized architectures; Church-Turing thesis; computational complexity*); Information theory (*reliability-error checking; efficiency-vs-speed of information; information asymmetry*); Game theory (*decentralized neural architecture; asymmetric information distribution*); Quantum brain—quantum computer; Artificial Intelligence (AI) and Artificial Life; Experimental and theoretical Neuroscience; brain evolution (*evo-devo*).

The historical development of the field, lead to the "obvious" realization that knowledge derived from Network Science (e.g., work by A-L. Barabasi, MEJ Newman, DJ Watts, and others) could contribute to understand how the brain works, interacts, manages task flexibly, and the underlying involvement of synaptic distribution, density and strength. Moreover, it has become clearer over time that network evolution could shed light on the evolutionary history of neural network architecture (and its governing principles; see for instance, Sterling and Laughlin, 2017). The subject has been treated from diverse points of view, derived from the application of different intellectual approaches (cellular neuroscience, computational modeling, connectomic analysis, philosophy of neurosciences, etc.). These different approaches suggested alternatives views of the roles of networks in the functionality of the brain. Most of them would deal with the general problem of representation, though more recent developments have changed the focus on the flow of information (the routing). In this context, as we will see in one of the SI papers, the contribution by D. Graham identifies the mode(s) of function of the internet network as a new frame of reference to understand (aspects of) brain function.

Another essential issue related to network evolution we wanted to address in this SI was the role of self-organization and complex systems in shaping brain (any brain) architecture and its evolution (e.g., works by I. Prigogine and G. Nicolis, C. G. Langton, S. Kauffman, S. Kelso, P. Bak, S. H. Strogatz, C. Gershenson and F. Heylighen, R. Solé, and many others). Concepts such as "emergent properties" or "organizational levels" come to mind as relevant here.

Finally, being well aware that the relationships between Brain and Computer encompass a vast spectrum of topics from Natural Sciences, Mathematics, Computer Science, Psychology and Philosophy, we needed to narrow our scope: some of the topics we opted not to deal with were Consciousness, Behavior, Language, and Culture.

Before dealing with the articles presented in this SI (see Table 1), and what they contribute to the debate, we revisit some critical, and necessary, concepts/topics: machine(s), metaphor and analogy in science, and brain(s). In the following text, the references belonging to this Special Issue are identified by a (*) as a superscript of the year of publication.

1.2. Semantics: Concepts and definitions

"I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'" (Turing, 1950). No doubt, Alan Turing had a clear understanding of the importance of semantics in this context. Likewise, all Authors and Editors of this Special Issue recognize semantics as a crucial concern in the brain-computer analogy debate.

Indeed, the authors identify a number of terms whose current definitions are "problematic," and need therefore to be taken with caution. First and foremost, is the definition of *"computer"* (Brette, 2022*; Richards and Lillicrap, 2022*). And the list continues with *"computing"* and *"recursion"* (Danchin and Fenton, 2022*), "*algorithm*" (Brette, 2022*; Richards and Lillicrap, 2022*; Roli et al., 2022*), "*computable function*" (Richards and Lillicrap, 2022*), "*robot*," *"program,"* and "*software*" (Bongard and Levin, 2021*), "*information*" (Cobb, 2021*; Gershenson, 2021*; Danchin and Fenton, 2022*), "*Artificial Intelligence*" (Roli et al., 2022*), *"intelligence"* (Gershenson, 2021*), "*cognition*" (Fraser et al., 2021*).

In the following, we are focusing on two specific, fundamental, issues: the definition of "*machine*," and the distinction between *"metaphor"* and "*analogy*." As for the former, in this Special Issue, Cobb mainly deals with images of the brain in history, and Bongard/Levin express their concern about an "outdated" view of term machine. In this paper, we follow the history of how "different kinds of machines" have best represented the brain. As for the latter, the distinction between metaphor and analogy is

TABLE 1 The articles in the Special Issue.

Article(*)	Authors	Title
0	*Matassi G. and Martinez P. (**)*	*The Brain-Computer Analogy – "a Special Issue"*
1	Cobb M.	A Brief History of Wires in the Brain
2	Gomez-Marin A.	Commentary: Metaphors We Live By
3	Chirimuuta M.	Artifacts and levels of abstraction
4	Brette R.	Brains as Computers: Metaphor, Analogy, Theory or Fact?
5	Bongard J. and Levin M.	Living Things Are Not (20th Century) Machines: Updating Mechanism Metaphors in Light of the Modern Science of Machine Behavior
6	Richards B. A. and Lillicrap T. P	The Brain-Computer Metaphor Debate Is Useless: A Matter of Semantics
7	Fraser P., Solé R. and De las Cuevas G.	Why Can the Brain (and Not a Computer) Make Sense of the Liar Paradox?
8	Roli A., Jaeger J., and Kauffman S. A.	How Organisms Come to Know the World: Fundamental Limits on Artificial General Intelligence
9	Danchin A. and Fenton A. A.	From Analog to Digital Computing: Is Homo sapiens' Brain on its Way to Become a Turing Machine?
10	Davis M.	The Brain-As-Computer Metaphor
11	Gershenson C.	Intelligence as Information Processing: Brains, Swarms, and Computers
12	Graham D.	Nine Insights From Internet Engineering That Help Us Understand Brain Network Communication

(*): The articles' numbering in this Special Issue, from 1 to 12, is the one also used in the main text, and in Figure 1. (**): this paper.

only touched on by Brette and Gomez-Marin. Consequently, we deem necessary to have a more detailed description of these concepts, which we deal with in turn, and offer a revised definition for the latter.

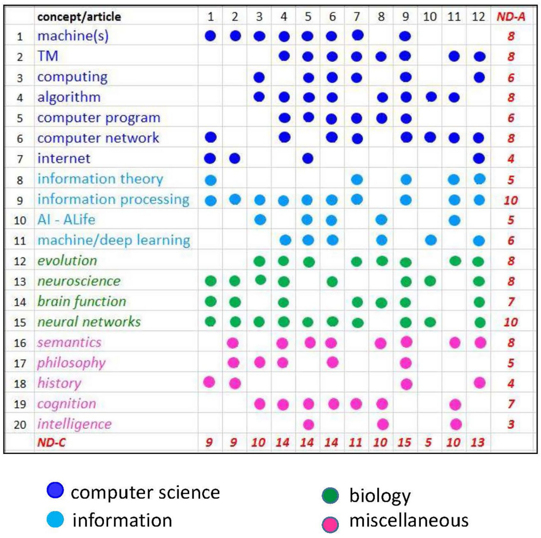
1.3. Machine(s)

Among neuroscientists there is a general opinion (sustained over decades of research) that what the brain "is" depends on how you study it. We live in a mechanical age, so we study it as a machine. In this context, we should question ourselves, upfront, what machines are and how our view of them has changed over time. In this SI, Cobb and Bongard/Levin introduce us to the ways we came to understand machines, in the past and nowadays. The underlying rationale for discussing "machines" is that the method of study has determined always what we have learned about them and how we have transferred these methodologies to study the brain.

For a long time, brains have been assimilated to certain kinds of "machines." The idea can be traced back, for a solid articulation, to the Cartesian view of the World, understanding machines as any physical system capable of performing certain functions. Descartes' body organs operate in purely mechanical fashion, and in this proposal, Descartes "creatively" adapted previous theories (Aristotle, Galen, etc.) to his own mechanistic program (Hatfield, 2012).

The form the machine analogy has taken over the years has suffered many transformations, adopting at every time the dominant mechanical view of the world (hydraulic, electrical or informational). In this SI, Cobb has revisited some of the historical views, with Bongard/Levin adding a perspective that includes recent developments in Artificial Intelligence. Interestingly, with the 20th century advancements in molecular biology, the machine analogy has been transferred from the whole tissue to the biochemical components that control its different functions. In this sense, the brain is equated to a soup/stew of highly coordinated chemical ingredients (molecular machines) that, ultimately, enable our rich psychological experiences. The whole field of neurochemistry, which foundations were laid in Europe, notably France and Germany, in the late 18th and early 19th centuries, with an important momentum gained in the 60 and 70's of the 20th century (Boullerne et al., 2020) is based on the assumption that interrelating chemistry and function in the nervous system is a most productive avenue to understand the brain (e.g., Brady et al., 2011).

From a functional perspective, over time our view of the brain has been transformed from a rather passive, fluid conducting device, to a more active, information processing one (a device able to compute; calculate in the original meaning). The computer (originally a person able to "compute" operations) was, and is in good part, understood as a mechanical device with certain properties (ability to store, retrieve, and process data). At this stage, it is relevant to consider that in spite that the *English word "computer" is meant to signify* (programmable machine that can store, retrieve, and process data; Encyclopedia Britannica) the different Romanic languages retain the original meaning of computers as a person who either organizes or computes datasets (e.g., *"ordinateur, ordenador," in French and Spanish*). In any case, we now universally use the term as meaning a device, usually electronic, that processes data according to a set of instructions. In this context, it is worth to remember that a more precise



	concept/article	1	2	3	4	5	6	7	8	9	10	11	12	ND-A
1	machine(s)	●	●	●	●	●	●	●		●				8
2	TM				●	●	●	●	●	●		●	●	8
3	computing			●		●	●	●		●			●	6
4	algorithm			●	●	●	●		●	●	●	●		8
5	computer program				●	●	●	●	●	●				6
6	computer network	●			●		●	●		●	●	●	●	8
7	internet	●	●			●							●	4
8	information theory	●						●		●		●	●	5
9	information processing	●	●	●	●	●	●	●		●		●	●	10
10	AI - ALife			●		●	●		●			●		5
11	machine/deep learning				●	●	●		●		●		●	6
12	*evolution*			●	●	●		●	●	●		●	●	8
13	*neuroscience*	●	●	●	●		●			●	●		●	8
14	*brain function*	●	●		●			●	●	●			●	7
15	*neural networks*	●	●	●	●	●	●	●		●	●		●	10
16	*semantics*		●		●	●	●		●	●		●	●	8
17	*philosophy*		●	●	●		●			●				5
18	*history*	●	●							●			●	4
19	*cognition*			●	●	●	●	●	●			●		7
20	*intelligence*					●			●			●		3
	ND-C	9	9	10	14	14	14	11	10	15	5	10	13	

● computer science
● information
● biology
● miscellaneous

FIGURE 1
The Special Issue at a glance. The figure is a kind of [0,1] matrix that describes the 12 articles in this Special Issue by means of 20 concepts (keywords) whose presence is denoted by a dot. The external column shows the number of Articles per concept (ND-A), the external row the number of Concepts per articles [ND-C; i.e., the Node Degree (ND), the number of links per node, in the corresponding bipartite graph, not shown]. Articles are as follows: Cobb (1), Gomez-Marin (2), Chirimuuta (3), Brette (4), Bongard-Levin (5), Richards-Lillicrap (6), Fraser (7), Roli-Jaeger-Kauffman (8), Danchin-Fenton (9), Davis (10), Gershenson (11), Graham (12). Abbreviations: TM-Turing machine; AI-Artificial Intelligence; ALife-Artificial Life.

characterization of the computer was given early on by Mahoney in his historical review of computing in which the computer is being specifically defined as a fundamentally tripartite structure, which reflect the contributions of three historical disciplines concerned with the nature of this "machine/device" (electrical engineering; computer science and software engineering). Moreover, Mahoney clearly stated what those contributions were in the summary sentence: *"between the mathematics that makes the device theoretically possible and the electronics that makes it practically feasible lies the programming that makes it intellectually, economically, and socially possible"* (Mahoney, 1988).

Interestingly, and as a result of the inception of the information age (in the 1940's), where information content and logical operations were introduced by logicians such as Alonzo Church and Alan Turing, the most salient analogy for the structure and function of the brain has been the computer, an instantiation of the so-called Turing machines (TM). The focus has changed from the instantiation of the machine to the underlying operative. In his seminal 1950 paper, Turing describes "machines" as those artifacts produced by "*… every kind of engineering technique*," and suggests to identify them with *"electronic computers"* or "*digital computers*," given the interest in his historical time in those devices (Turing, 1950). He gives a definition of computer as a finite state machine (a mathematical model of computation). An extended quote from Turing seems appropriate here.

*"A digital computer can usually be regarded as consisting of three parts: (i) Store [of information] … corresponding to the paper used by a human computer … [and] … the book of rules ", (ii) Executive unit [carries out calculations], (iii) Control [handles the correct use of instructions]. … digital computers … fall within the class of discrete state machines. … This machine could be described abstractly as follows. The internal state of the machine (which is described by the position of the wheel) may be q1, q2 or q3. There is an input signal i0 or i1 (position of lever). The internal state at any moment is determined by the last state and input signal according to the table [of instructions]. … These are the machines which move by sudden jumps or clicks from one quite definite state to another. … the digital computer … must be programmed afresh for each new machine which it is desired to mimic. This special property of digital computers … is described by saying that they are universal machines."*Needless to say, not all brain-computer metaphors require traditional TM or von Neumann architectures. We now have parallel or quantum computing, for instance, and these modalities have enriched our view of what computers can do (see Kerskens and Lopez-Perez, 2022, suggesting that our brains use quantum computation). However, one particularly persistent (and relevant here) view of computing and brain functions emphasizes the parallel architectures that both utilize, breaking up problems into smaller units that are executed by different components, all communicating through a shared memory. Thus, when comparing computers and brains, a common inference is that both systems, essentially, rely on parallel processors. This is not an accurate representation of the similarities, and for a number of reasons. (i) Brains and computers use different orders of magnitude (6 or 7) of independent (computing) units. (ii) While processors in a computer are "all purpose," the human brain has specific areas specialized in processing different kinds of input. (iii) There are big differences in reliability and adaptability between brains and computers (a concept linked to that of "reprogrammability" in both systems), where brains information-processing systems are intrinsically "noisy" (Faisal et al., 2008) and this explains the differences in reliability and adaptability between them and the computers. (iv) Brains are fast at recognizing patterns from complex data, which (in many cases) are not possible by massive parallel computing systems (Hawkins and Blakeslee, 2004). These factors seem to suggest that parallel processing in the brain is never "truly" parallel and that reprograming in brains and computers rely on different network "reconfiguration" strategies (re-routing in machines versus neural plasticity in living systems). As the needs arise (e.g., "landscape modifications"), the adaptability of biological systems (e.g., brains) is a unique property derived from the plasticity of cell and circuit configurations, and a result of both genetic and epigenetic factors controlling birth, death and connectivity of brain neuronal sets.

In the following, we return to the processing system and provide an "accessible" description of a Turing machine (TM) that should be useful to understand the metaphor used for the brain. Briefly, a TM, or an *"automatic machine"* as Turing called it (Turing, 1937), is an abstract idealized model of a simple kind of digital computer. The machine input is a string of symbols each one of them carried by a single cell on a linear tape. The machine possesses some sort of read-write scanning head that considers one cell at a time. It is an automatic machine (i.e., at any given moment, its behavior is completely determined by the current state and symbol, the "*configuration*," being scanned). It is a machine capable of a finite set of configurations. A finite set of rules (i.e., the *program* representing the *algorithm*) instructs the machine what to do in response to each symbol (i.e., erase, write, move left, move right, do not move). In principle, for any function that is computable (i.e., a function whose values may be computed by means of an algorithm), there is a TM capable of computing it. This logically implies the property of imitating another machine, meaning that there is a Universal Turing Machine (UTM) capable of simulating any other TM performing different tasks, by reading the corresponding set of rules from the tape. This is the theoretical model of a *programmable computer* (for more details on TM-UTM, see Gershenson, 2021*; Danchin and Fenton, 2022*; Richards and Lillicrap, 2022*).

Given the definition of a TM, it soon became clear that the brain (or mind) could be equated to a computational system very similar to a TM, and with many of the mental processes very similar to computations performed by a TM. Some authors consider that this identification of brains with TM is too strong, and thus, an adherence to it is called the "hard position."

This "hard position" is being criticized by other authors saying that neither the principles, nor the materials or the way they are utilized (or organized) in a brain can be equated to a TM, except, perhaps, in the way both perform arithmetic operations (some authors deem the whole comparison "vacuous"; see Epstein, 2016). We are not going to delve on this problematic here, just want to stress the enormous influence that Turing machines have had, as computational neuroscientists have maintained (not all), for the last 80 years, in the view of brains as computer (working as a TM). This model was vindicated early on when neuroscientists realized that neurons were performing their physiological roles, firing action potentials, in a "all or none" fashion. This view was mostly promoted by Warren McCulloch and Walter Pitts in 1943, who also saw neural circuits in the brain as circuits of logical gates. Modern neuroscience has revealed more complex firing patterns, as well as complex patterns of firing regulation, adding nuances to the original McCulloch-Pitts view.

In the comparison between computers and brains, the semantics issue has often been raised, in one form or another. In particular, in the 80's John Searle asked the question: "*Can a machine ever be truly called intelligent?*" (Searle, 1984). The question was encapsulated in the well-known "*Chinese room*" argument. It suggests that however well one programs a computer, nonetheless the machine does not understand Chinese; it only simulates that knowledge, and therefore this behavior cannot be equated with intelligence. Searle argues that his thought-experiment underscores the fact that computers merely use syntactic rules to manipulate strings of symbols, but have no understanding of their meaning. The issue of "meaning" is not further explored here, though we recognize its enormous interest. Searle's main conclusion was that passing the "Turing Test" is inadequate as an answer (see also Cole, 2020).

All in all, in spite of the historical fortunes (and misfortunes) of the brain-computer analogy, the use of this *"metaphor"* is widespread, a testimony of which can be found in the different papers of this special issue. The subject remains fertile and open for further discussions.

1.4. Metaphor and analogy in science

The definitions of metaphor and analogy, at least in English dictionaries and encyclopedias, serve well to illustrate how muddled these concepts still are, in spite of the massive literature, in both Science and Humanities, devoted to them. A telling example comes from the Merriam-Webster in which metaphor **is** defined as "*a figure of speech in which a word or phrase literally denoting one kind of object or idea is used in place of another to suggest a likeness or analogy between them.*"

The definition of the two concepts has been "adapted" in different branches of human knowledge. Here we are only concerned with the meaning(s) of Metaphor and Analogy (M&A) in (western) scientific thought. As we will see in the following sections, in Science, and also in this SI, M&A are used as synonyms, yet they are not. Aware of this, in the title of this Research Topic (*Current thoughts on the Brain-Computer analogy—All metaphors are wrong but some are useful*) we intentionally, and provokingly, used metaphor and analogy as synonyms; and it is precisely in Science that the distinction is more conspicuous. This is the rationale for discussing in detail the issue in this section.

It is important to note, upfront, that some argue that metaphor and analogy have actually no place in science (for a discussion see Haack, 2019; Reynolds, 2022), though others claim that M&A are essential for scientific creativity (a position sustained in this article). According to Ziman "… *scientific theories are unavoidably metaphorical*" (Ziman, 2000), and it has been suggested that they are "*the basis of our ability to extend the boundaries of human knowledge*" (Yohan, 2012). Moreover, the aptitude for metaphorical and analogical reasoning is an essential part of human cognition. Undeniably, M&A have been a powerful way to communicate knowledge and consequently a powerful tool in education and learning. Just think of how many times we use metaphorical language to convey concepts to students in our own teaching experience (Kovac, 2003). Scientific M&A can guide scientific discovery, hypothesis and theory, and plays also an important role in adapting scientific language to the world. As Kuhn put it "*Metaphors play an essential role in establishing links between scientific language and the world*" but what is crucial (see also section 3) is that "*… Those links are not, however, given once and for all*" (Kuhn, 1993). Then choosing the "right" metaphor may be regarded as part of scientists/teacher work and ultimately becomes a form of art (Haack, 2019). In the following, we discuss M&A in more detail, given the relevance we assign to them in the context of discussing our current images of the brain.

1.5. Metaphor

The literature on metaphor is overwhelming and definitions abound. Robert R. Hoffman reasons that scientific metaphors appear in a variety of different forms and serve a variety of functions, and it makes a rather exhaustive list of them (Hoffman, 1985). One example for all, the *"Tree of Knowledge."* In its various flavors over the centuries, it is certainly one of the founding metaphors of human civilization, not only of Science (Lima, 2014). And, to an evolutionary biologist (like the two authors of this paper), there is hardly a more fundamental metaphor than the *Tree of Life,* which Darwin, and Lamarck (1809) before him, used to illustrate his theory of descent with modification and depicted in his *"Diagram of diverging taxa"* (Darwin, 1859, pp. 116–117 in 6th edn). Indeed, the tree metaphor has been used in evolutionary biology ever since. However, and most notably, based on the regained awareness of the evolutionary impact of the phenomenon of gene flow between species (a.k.a. Horizontal/Lateral Gene Transfer) less than three decades ago, a new metaphor has emerged to account for the diversity of species: The *Network of Life* (Doolittle, 1999; Martin, 1999; Ragan, 2009). Incidentally, as an historical note, contrary to common knowledge, the network metaphor predates that of the branching tree. Indeed, it is dated 1750 and credited to Vitaliano Donati, whereas the first use of the tree metaphor is attributed to Pallas in 1776 (cited in Ragan, 2009). The example of the use of Trees and Networks in evolutionary biology is mentioned here specifically to emphasize that the two metaphors are complementary; we consider this position as pivotal in our review essay for both trees and networks have been specifically used in modeling our ideas of the brain and its evolution (see also section 3).

As to the definitions of metaphors in science, for the sake of brevity, we single out two of them (JC Maxwell, and Lakoff and Johnson) adding three illustrative examples for a better understanding.

James Clerk Maxwell wrote "*The figure of speech or of thought by which we transfer the language and ideas of a familiar science to one with which we are less acquainted may be called Scientific Metaphor*" (Maxwell, 1870). This would refer to a "logical semantics" view of metaphors, very much used in science and everyday life.

However, and in the classical definition by Lakoff and Johnson, the concept of "mapping" is also introduced. This leads them to state that "*The essence of metaphor is understanding and experiencing one kind of thing in terms of another*" (Lakoff and Johnson, 1980). And also introduce the different idea that the "*Metaphor is the main mechanism through which we comprehend abstract concepts and perform abstract reasoning… Metaphors are mappings across conceptual domains*" (Lakoff, 1993a). Hence, metaphors "become" conceptual tools (aids in understanding). Therefore, as Humar put it "*A metaphor links two domains by mapping attributes from one onto the other. Thus, metaphor is an act of transferring… [where]… the key terms, 'target' and 'source', were introduced by Lakoff and Johnson … For instance, the biological metaphor 'genes are text' links the source 'text' and the target 'genes'*" (Humar, 2021). Black (1962) points out, in this context, how similar is the "standardized" Oxford English Dictionary (OED)[1] description of metaphor to the one described above: "*The figure of speech in which a name or descriptive term is transferred to some object different from, but analogous to, that to which it is properly applicable; an instance of this, a metaphorical expression*".

Interestingly, the etymology of the term "metaphor" originates from the ancient Greek noun "*metaphora*" (μεταφορά), which is derived from the verb "*metapherein*" (μεταφέρειν), originally meaning "to transfer," "to transform." Or else, derived from *μετα* (*over, beyond*) and *πηερειν* (*to carry*). It all depends on what we mean by "transfer" or "carry beyond" in the above definitions; more precisely we may ask: What is being transferred?

Before delving into the next relevant concept of analogy, we need consider another rather problematic term which is very often linked or likened to metaphor: the concept of "model." Often, in the scientific literature there is no clear distinction between model and metaphor. In fact, we think that a distinction needs be made, for clarity. Contrary to a metaphor, a model (a conceptual model) has for us a narrower scope and, being a hypothetical representation of a system, it aims at simulating and understanding reality (e.g., a biological model; see also Ziman, 2000). Moreover, "*… a model is, in its etymological and technical sense, a substantive thing which is the best or ideal representative of something else. All other uses of the word "model" are metaphorical extensions of this basic meaning*" (Hoffman, 1985). Therefore, we see models as methods or representations aimed to understand, and predict, specific patterns.

To complete this section, we would like to propose a "concept" of metaphor that does not require, but accepts, the use of the mapping concept (but see section 1.6, below). The type of metaphor we have in mind is founded on a visual perception. It is the visual image that is the driver for scientific insight and provides educational power. As an example, we identify three metaphors that best illustrate this idea: Adaptive Landscapes by Wright (1931), Epigenetic Landscapes by Waddington (1957), and the Gene Regulatory Network by Davidson and Peter (2015). For a recent, and more extensive, discussion of the use of metaphors in science, with its dangers and pitfalls, we refer to the excellent new book by Reynolds (2022).

1.6. Analogy

Metaphors may be a source for "*analogies*" (and *"similarities"*) and may guide building models. Among the definitions of analogy

[1] https://www.oed.com/view/Entry/117328?redirectedFrom=metaphor&

given in the OED there are the following: (a) *A comparison between one thing and another, typically for the purpose of explanation or clarification; (b) Biology: The resemblance of function between organs that have a different evolutionary origin*. Our focus here is on the first, more general, definition.

Atran (1990) traces back the concept to Aristotle and his effort to compare structures and functions between man, other animals and plants; "*… Aristotelian life-forms are distinguished and related through possession of analogous organs of the same essential functions*." Along the same line, the concept of "*analogue*" was introduced in comparative anatomy in 1843 and defined as "*a part or organ in one animal which has the same function as another part or organ in a different animal*" (Owen, 1843, p. 374). Atran brings us to a more generalized version of the analogy concept, also mentions the Newton's concept of "Analogy of Nature" (*ibid* p. 232) and points out that this analogy "*… combines two older ideas: the theological "Chain of Being" through which Nature seeks Divine Perfection, and the unity of causal pattern in the macrocosm and the microcosm*."

But it is in her classic book that Mary Hesse describes in considerable detail both scientific models and analogies (Hesse, 1970). A dialog is imagined between two men of science: *Campbellian*, who argues that analogies and "*models in some sense are essential to the logic of scientific theories*" and *Duhemist*, who denies it. *Campbellian* identifies three types of analogies: positive, negative and neutral. Two physical objects or systems have *positive analogy* based on their shared "properties": "*Take, for example, the earth and the moon. Both are large, solid, opaque, spherical bodies, receiving heat and light from the sun, …*" yet, the same objects may differ in a number of respects: "*On the other hand, the moon is smaller than the earth, more volcanic, and has no atmosphere and no water … there is negative analogy between them*." *Neutral analogies* are "properties of the model about which we do not yet know whether they are positive or negative analogies." Note that *Campbellian* too is concerned with semantics: "*But first let us agree on the sense in which we are using the word model*." Thus Hesse tells us that analogies can have specific "values": positive, negative or neutral.

Humar has posed a dichotomy between "structural metaphors," such as those described above and functional ones. In fact, "*functional metaphors … draw attention to a similarity in function between a source and a target are also found in ancient scientific literature*" (2021). And again, Gentner and Jezioreski (1993) contend that an underlying idea pervades the use of any concept of analogy "*The central idea is that an analogy is a mapping of knowledge from one domain (the base) into another (the target) such that a system of relations that holds among the base objects also holds among the target objects. In interpreting an analogy, people seek to put the objects of the base in one-to-one correspondence with the objects of the target so as to obtain the maximal structural match.*" More so, broadly speaking, Hoffman sees the distinction between the two concepts as a chicken-and-egg problem, analogy regarded as the "*psychological egg*" and metaphor the "*chicken*" (Hoffman, 1985, p. 348).

Finally, we suggest the use of two criteria, the structural and the functional in the very definition of analogy (in science), and indicate the latter as its most characterizing property. In doing so, we do link this definition of analogy with the Lakoff-Johnson definition of metaphor and its associated action of "transfer." We think this definition of analogy is more pertinent (of practical importance) here because metaphors are not intended to provide a solution to a given problem, they have no explanatory power. In contrast, analogies do have an explanatory power, and enable to make connections to understand the structure/function of a given system based on the knowledge acquired on another system. A telling example for the "explanatory role" of analogy is the transfer of *themata* [*sensu* Holton, in Ziman (2000)] between different disciplines—for example, the notion of a "code" from information theory to molecular genetics.

1.7. Brain(s)

A key concept in this Special Issue is obviously that of a Brain, but how to define one? The brain, defined in simple terms, according to the Encyclopedia Britannica is: "*the mass of nerve tissue in the anterior end of an organism*." The brain integrates sensory information and directs motor responses. While this mostly represent the vertebrate condition, the substitution of nerves by neurons would be still a valid assertion. Brains as centralized structures are old, dating back to the origin of bilaterian animals in the Ediacaran Period (571 to 539 million years ago; Martinez and Sprecher, 2020). The coalescence of neurons in a pole of the larvae/animal allows a better, centralized, coordination of functions, and in that sense, brains have been also equated to "central processing units" (CPUs). How centralization has happened and the conditions that drove their appearance have been discussed before (see Martinez and Sprecher, 2020) and do not need a further discussion here.

Our ideas of the brain have changed radically over the centuries, mainly due to the lack of proper understanding of their physical constituents and the modes of functioning. Explanations have used the current metaphors that conformed the mechanical world at every age (see Cobb's historical account in this SI). Most recently, and with the instantiation of computing devices and the rise of the information age, computing and information processing have been our reference mark when thinking about brains and their activities. The current view originated early in the 20th century, when the brain tissue was systematically analyzed under the microscope. The presence of isolated cells organized as neural nets contributed to the view of the brain as a "machine" dedicated to compute and process information.

The intricate nature of brain connections (neurons and substructures) suggested the possibility that the brain is actually a connected set of wires, with complex architectures (see Cobb, 2020). Moreover, the discovery of chemical and electrical connections between neurons reinforced the image of a giant electrical device with multiple, complex, switching mechanisms.

It is the emergence of the information age, with the first devices able to "compute" operations, that led to a new model of the brain, which in addition to conducting electrical impulses, was assimilated to a complex computing device.

The integrative model of the neurons, with empirical data and modeling processes, was developed by pioneer cyberneticians/neurophysiologists Warren McCulloch and mathematician Walter Pitts (among others). McCulloch's brand of cybernetics used logic and mathematics to develop models of neural networks that embodied the functioning of the brain in the workings of the brain (Pitts and McCulloch, 1947). How accurate is this model? The question has been a subject of intense debate, to which some of the papers in this issue refer (e.g., Davis, 2021*; Fraser et al., 2021*). Ideas about how the information is processed, the speed of neuronal communication, the role of the neuron in integrating inputs, the routing of information and the correlation between firing patterns and brain activities (i.e., mental activities), have all contributed to the debate on the validity of using "computer" metaphors for understanding different aspects of neuroscience. The debate is alive today as it ever was.

At the base of our utilization of metaphors for specific organ systems is the consideration that the activities of the organ as properties "define" the realm (domain) and the contents of the metaphors. In this sense, brains are equated to computers because, at least according to some authors, they are actually performing "computing" operations (see Chirimuuta, 2022*). However, there is not a unified agreement on the use of this metaphor (others are explored in this SI by Gomez-Marin and Graham), and this has led to a heated debate on the meaningfulness of using some specific metaphors in neuroscience (see a later discussion in this paper).

One of the key issues discussed by many authors interested in modeling the brain and its functions revolves around the nature of information flow and how input signals are transformed into output behaviors, including the routing problem (see Graham, 2022*; in preparation)[2]. This is linked to the idea that our brain does not function as a linear processor in which the flowing streams of information, from input data to output realization (behavior) are not unidirectional, a "one-way street." Instead, many authors consider that the output of the brain's processing is the result of some "emergent properties" not linearly derived from the original inputs, properties that are not "just" the result of simple operations (addition/subtraction) of inputs. Some of the problems not solved by the different physical models of the brain are linked to the capacities for self-reference in human brains, or more generally the awareness of our own existence (consciousness). These problems are not easily dispatched by models of emergence, and a proof of the complexity that self-reference models have in computer science is shown by Fraser et al. in this SI. Once more, mathematical descriptions and observable reality are not easy to compare.

2. The 12 articles in the special issue

In this section, we present our own summaries of the 12 articles (see Table 1) in this Special Issue (SI), each of which (but one) has been endorsed by the corresponding author(s).

In Figure 1 we propose a graphical picture, a sort of a snapshot of the entire Special Issue, based on 20 concepts (keywords) we have arbitrarily selected. It is a kind of [0,1] matrix in which the presence of those concepts in a given paper is denoted by a dot. The usefulness of such a representation is self-explanatory. In the following, article summaries are listed by Authors' names (and number in Figure 1).

We have chosen as the opening article of this Special Issue Matthew Cobb's historical account of the metaphors used over the centuries to describe the brain and try to understand its functioning.

2.1. Cobb (1)

In the opening article of this Special Issue, Matthew Cobb, the author of the excellent book *"The Idea of the Brain"* (2020), provides a detailed and instructive history of the "*wiring diagram*" metaphor of the brain and explores its role, together with that of its associated metaphors, on the conception of the brain over the last two centuries.

His historical account of the use of metaphors for brain functioning starts in the 18th century stemming from mechanics, and the discovery of electricity (telegraph). Cobb identifies a drastic shift toward the end of 19th century with the appearance of " *… the telephone exchange, where messages can be flexibly routed.*" In the 20th century" … *two kinds of wiring diagram – that of the animal body and that of the computer – entered into dialogue*" (McCulloch and Pitts, 1943; von Neumann, 1958).

In the 21st century, the connectomic projects, which are aimed at a complete description of the structural connectivity of the central nervous system, became prominent, in many respects. Cobb criticizes these approaches mainly because they produce a static representations of the nervous system. He thinks that we should proceed from small circuits (controlling specific behaviors) to the whole map of neuronal connections, and gives the example of the lobster's stomach, whose processes are controlled by a few neurons, which has been studied (excruciatingly) for a long time, and for which we still do not have a full understanding. Moreover, quite rightly, Cobb points out that knowing the genome sequence cannot by itself explain the "functioning" of the corresponding

[2] Graham, D. (2022). Nine insights from internet engineering that help us understand brain network communication. *Front. Comput. Sci.* (in preparation).

organism, likewise " ... *the wiring diagram itself could not explain the workings of the human mind*."

Cobb mentions what is regarded as the most recent metaphor for brain function *"cloud computing or the internet*." On the one hand he acknowledges that " ... *it embodies plasticity and distributed function into our conception of the brain*," on the other hand, he points out its limits if the notion of robustness is taken into account" ... *the internet is designed to function even if key parts it removed, whereas some aspects of brain function can be decisively disrupted if particular areas are damaged*." Interestingly, the internet metaphor will be explored in great detail by David Graham, in the last contribution of this Special Issue.

As a cautionary note, Cobb warns us about the limits of the use of these metaphors to study brain function, mainly " ... *because of the plasticity and distributed function of most nervous systems*." The notion of brain plasticity is central to this Special Issue and alludes to the plasticity that nervous systems shown in individuals during their lives and linked to the processes of learning and memory acquisition.

The next three articles deal with conceptual issues, and are written by A. Gomez-Marin, M. Chirimuuta, and R. Brette. They tackle the problem of whether there is any foundation for the comparison between brains and computers. In different ways, they do that by interrogating the interrelated questions of what is a computer, how it can be characterized and the limitations that these characterizations, and their associated metaphors, have in our current understanding of both brains and computers.

2.2. Gomez-Marin (2)

Gomez-Marin introduces us to the well-known book "*Metaphors we live by*" (1980), authored by the cognitive scientists George P. Lakoff and Mark Johnson. In this seminal work, the authors provide a detailed analysis of the nature of metaphors, suggesting that metaphors, which were once known as mere "linguistic devices" (semantics), are mostly "conceptual constructions" that shape the way we think and act. In a sense, as Gomez Marin points out, the semantic role for metaphors is secondary to their conceptual (cognitive) nature. Following Lakoff (1993b), metaphor mapping (from one conceptual domain to another) would occur independently of their linguistic expressions, so there is a priority status given to their cognitive function, over those expressed in language terms. Or put it another way: the conceptual structure of metaphor is given more weight than the structure of metaphoric language.

In this context, Gomez-Marin revisits the analogy of computers and brains. After a brief historical overview of the most pervasive ways in which brains and computers have been visualized, Gomez-Marin draws our attention to the lesser-known images of the brain such as holograms and radio sets. The latter suggests the intriguing possibility that "brains would not create thoughts but receive and filter them."

Gomez-Marin summarizes his appraisal of metaphors with the advice that we apply them as pragmatic tools with the proviso that we should be always vigilant to avoid what he calls falling into a "metaphorical monoculture", which would become "a burden" rather than "a blessing."

2.3. Chirimuuta (3)

In a suggestive parallel, Chirimuuta comments on the assertion by different authors that the brain and computers (or any other complex artifact) could be made tractable by using multi-level approaches. These approaches use top down, functional characterizations of systems to compliment bottom up reductionist strategies. The important assumption is that the brain decomposes into relatively autonomous levels of organization, similar to the hardware-software distinction in computing.

However, as appealing the simile can be, Chirimuuta contends that several limitations need to be accounted for. (1) Low-level components (neurons in the brain) are not mere "hardware implementors" in brain information processing. In computers, the elements maintaining the physical integrity of the machine and the components performing information processing are different. Whether this separation occurs in the brain is far from clear. (2) Computers and artifacts are assembled differently. While computers are designed to ensure that high level functionality is relatively independent of variations in hardware the functionality of the brain may well depend on low level details often assumed to be irrelevant to cognition. Interestingly, the two alternatives are, again, assumed to be the products of two constructive methods: engineering, in the case of computers/artifacts, and evolution, in the case of brains. Chirimuuta, however, is concerned about oversimplifying the principles that govern the construction and functionality of complex biological systems, such as the brain.

2.4. Brette (4)

The question central to Brette and Richards/Lillicrap (see below) is a semantic one, they ask "*What is a computer?*" Brette states that *both in common and technical usage a "computer" is thought of as a "programmable machine*." Then, while pointing out that in computer science there is no formal definition of computer, he draws our attention on the concept of "*program*" defined as " ... *a set of explicit instructions that fully specify the behavior of the system in advance ("pro-", before; "-gram", write)*."

Moreover, and quite appropriately, Brette discuss the notions of algorithm and computation in the brain. At this point, two deep questions, both from evolutionary and philosophical perspectives: "*what is a brain program*"? and "*who gets to "program" the brain?*" The reasoning those entail, seem to lead to a logical consequence " ... *The brain* *might not be a computer, because it is not literally programmable*." Offering a definition of metaphor and analogy, Brette concludes that the brain-computer metaphor seems to be of little use, if not misleading, for it provides, according to the author, a reductionist view of cognition and behavior. This conclusion contrasts sharply with that of Richards and Lillicrap.

The next contributions deal with different problematics arising from the brain-computer comparisons; whether these are semantic misunderstandings (formal definitions in the field) or with misleading assumptions of what a computer or a brain can do. The papers by Bongard and Levin and Richards and Lillicrap deal with a fundamental problem that affect all definitions. The definitions of concepts bear very much the stamp of the fields in which they are generated (e.g., computer science, engineering or neurobiology). This straightjacket affects the way we conceive the possibilities of what a computer or a brain can do. Revised versions of those concepts should liberate the concepts from the "semantic constrains" that those fields have imposed in them. Here, brain, computer and machines are the three examples analyzed in detail. In the following three papers, authored by Fraser et al., Roli et al., and Danchin and Fenton the subject of software (the running of algorithms) and how brains and computers deal with processing information is clearly put. All authors discuss the idea of to what extent Artificial Intelligence should be able to reproduce behaviors of living organisms. Irrespective of the general optimism in the possibilities of Artificial Intelligence, these authors introduce some cautionary notes; which cast some doubts on the real possibilities of "*imitating*," for instance, human behaviors. "Agency" and "self-reference" become clear stumbling blocks. One last paper in this section, authored by Davis, suggests a series of questions posed by the analogy, asking himself (and the people in the field) to what extend they have been answered and what the answers would add to the debate.

2.5. Bongard and Levin (5)

It has been assumed for a long time that life and machines are fundamentally different entities, and that the former can't be reduced to the latter (see Nicholson, 2013). Bongard and Levin contend that this dichotomy is mostly based on an old conception of machine, a 17th to 19th century vision that doesn't account for modern development in disciplines such as Artificial Intelligence, Bioengineering, etc. In this context the authors re-visit the problem and ask: "does a suitable machine metaphor apply sufficiently to biology to facilitate experimental and conceptual progress?." The path toward understanding this goes from a clear definition of what a machine is, and the properties characterize them, to a critical appraisal of what modern science and technology tells us about those properties. Do these properties are clearly demarcated between alive (or evolved) and engineered "things"? In view of modern developments in the above-mentioned sciences it becomes harder and harder to sustain a clear separation between these two "systems", with borders becoming more fluid as modern engineering progresses. The authors emphasize the fact that the analysis of properties once associated to the living beings in newly developed machines clearly show that the boundaries between those, once considered unmistakably different systems, are nowadays becoming blurred. Several, and very detailed, examples are provided. At the end they try to provide a new working definition of machine that can accommodate all of our newly gained insights.

2.6. Richards and Lillicrap (6)

Richards and Lillicrap emphasize the fact the word "computer" is given different definitions in different disciplines, and specifically they contrast the definition used in computer science with the one used by academics outside computer science. According to this argument, much of the debate about the brain-computer metaphor would be just a matter of semantic disagreement. End of the debate. Is it so?

While the common usage of "computer" is straightforward—"*human-made devices (laptops, smartphones, etc) that engage in sequential processing of inputs to produce outputs*"—this is certainly not so for the notion in computer science. The authors carry out an in depth analysis of the notion of "*computer*" in computer science, definition based on two other notions, those of "*algorithm*" and "*computable function*." An "*algorithm*" can be informally defined as a sequence of finite logical steps that mechanically lead to the solving of a problem. A "*computable function*" is "*any function whose values can be determined using an algorithm*." The formal definition of algorithm was developed independently by mathematicians Alan Turing and Alonzo Church in 1936–19837, and the authors introduce the Church-Turing Thesis.

The authors provide a formal definition of "*computer*"*: a "physical machinery that can implement algorithms in order to solve computable functions*." They stress that this definition "*is important because it underpins work in computational neuroscience and AI*." The authors also describe the applications of this definition to brains and discuss its limits.

In sum, Richards and Lillicrap invite us to contrast the two definition of "computer," inside and outside computer science. If one adopts the definition from computer science "*one can … simply ask, what type of computer is the brain?*" However, " … *if one adopts the definition from outside of computer science then brains are not computers, and arguably, computers are a very poor metaphor for brains*." End of the debate?

2.7. Fraser et al. (7)

One of the properties associated to (human) brains is the capacity of being conscious of its own existence. This seems an obvious difference to any other standard machine, including computers. Fraser and collaborators (see Fraser et al., this issue) make a good case for this, by appealing to self-reference statements, which cannot be resolved by computers. The case is clearly stated: verbal statements like "the sentence presently being uttered is false" are being "understood" by our brains. A computer presented with it, however, enters into an "endless loop," with no resolution. How do brains deal with the above paradox? Fraser et al. present an elegant dynamical model, in which brains are composed of interacting units (modules) moving through time (the strange loop model). The model suggests a way the brain has for dealing/resolving the paradox, and this is by extending the analysis of the inconsistency over time (deconvoluting it along this axis). Temporalizing the problem, the brain is able to cope with the paradox. This avoids the system (the brain) to enter into the endless loops that would characterize the response of a computer.

2.8. Roli et al. (8)

Roli, Jaeger and Kauffman put the focus on the fields of Artificial Intelligence (AI) and Artificial Life (Alife). The notions of Natural Intelligence and Artificial General Intelligence (AGI) are contrasted, the latter being defined as "*the ability of combining 'analytic, creative and practical intelligence'*." The ultimate goal of AI and ALife would be to create a computational or mechanical system (an AI-ALife-agent; e.g., a robot) able to autonomously (i.e., without human intervention) identify, appraise and exploit new alternative opportunities (dubbed *affordances*) so that to evolve and innovate in ways equivalent to a natural organism (autonomous Bio-agent). Affordances are here defined as "*opportunities or impediments on [a] path to attain a goal*." The authors argue that AI-and ALife-agents cannot "*evolve and innovate in ways equivalent to natural evolution*" for current AI algorithms do not allow such capability (broadly dubbed *"agency"*) since they cannot transcend their predefined space of possibilities (determined by the human designer). Moreover, they show that the term *"agency"* refers to radically different notions in biology and AI research.

As possible objections to their position, the authors mention (i) deep-learning algorithms and (ii) unpredictability of AI systems (e.g., playing chess, composing music); they address and dismiss both.

Finally, citing the work of William Byers and Roger Penrose, the authors distinguish the capabilities of Natural and Artificial intelligence using the notion of creativity in mathematics, creativity "*which does not come out of algorithmic thought but via insight*," which is not formal and involves shifting frames.

2.9. Danchin and Fenton (9)

Danchin and Fenton's paper correlates notions from neurobiology and computer science. Parallels are also drawn between computing, genomics and species evolution. The focal point of the paper, made explicit in the title, is built around the concept of Turing Machine (TM); an entire section of the paper is devoted to it. Noteworthy, the TM concept is also transposed into biology.

In order to explore the potential analogies between brain and computer, an issue of semantics is addressed first: the definition of computing. The differences between analog and digital computing are contrasted, with emphasis on the fact that analog computation implements of a variety of feedback and feedforward loops, whereas digital algorithms make use of recursive processes. Recursion is a central concept in this essay; it is a characteristic feature of the digital world of a TM. Recursion allows one of the steps of a procedure (e.g., set of rules of the TM machine) to invoke the procedure itself. "*A mechanical device is usually both deterministic and predictable, while computation involving recursion is deterministic but not necessarily predictable.*" The brain does certainly some sort of computation, and "*with remarkable efficiency, but this calculation is based on a network organisation made up of cells organised in superimposed layers, which gives particular importance to the surface/volume ratio … This computation belongs to the family of analog computation*" (A. Danchin).

In an extremely useful Table, the key features of a Turing Machine, a digital computer, and the human brain are compared. The authors conclude that brains are not digital computers. However, they speculate that the recent (in evolutionary time) invention of language in human history, and writing in particular (maybe around 6,000 years ago), might constitute a step toward the evolution of "*the brain into a genuine (slow) Turing machine*."

2.10. Davis (10)

Davis' short paper suggest a range of questions that bear into our use of brain computer analogies. His focus is on the programing of brain processes. What kind of algorithms use the brain to navigate the world? Would computers be able to simulate those? Davis suggest that modern use of optimization algorithms (network training) should provide an avenue, improving over the longer (older) numerical computations. He ends up by posing a provocative hypothesis: consciousness could play the role of an "interface to the brain's operating system." Definitely, questions that still remain unresolved in the fields of computer and Artificial Intelligence.

The paper by Carlos Gershenson introduces a different perspective to this SI by bringing to the fore the notion of intelligence, and collective (swarm) intelligence in particular and by linking it to the theory of information processing.

2.11. Gershenson (11)

The main take of the MS is stated in its Title: the focus is on Intelligence that is studied in terms of information processing. This approach could be applied to brains (single and collective), and machines.

A major issue arises immediately: There is no agreed definition of intelligence; semantics again. Gershenson defines intelligence in terms of information processing: "*An agent a can be described as intelligent if it transforms information … to increase its 'satisfaction' … Examples of goals are sustainability, survival, happiness, power, control, and understanding*." In previous work, Gershenson suggested to use measures of information as a tool to study complexity, emergence, self-organization, homeostasis, and autopoiesis (Fernandez et al., 2014); here he aims to extend this approach to cognitive systems, including brains and computers.

Information, a new semantic challenge. Gershenson presents a definition of information quoting the classic work of Shannon. Our attention is drawn on the meaning of the message being transferred; in this context, the failure of Laplace daemon (and Leibniz mill for that matter) is instrumental in identifying a crucial, and much-overlooked notion (not only) in Biology: the existence of different scales, different frames of reference, which (ought to) modify the models and hypothesis for a given phenomenon. "*Even with full information of the states of the components of a system, prediction is limited because interactions generate novel information*."

A stimulating comparison is offered between the intelligence of "the single brain" and the collective intelligence of swarms (groups of humans, animals, machines). In the case of insect swarms, which can be described as information processing systems, the processing is distributed. Gershenson compares the cognitive architectures of brains and swarms, and identifies a key feature distinguishing the two: "*the speed and scalability of information processing of brains is much superior than that of swarms: neurons can interact in the scale of milliseconds, … insects interact in the scale of seconds, … [in practice,] this limits considerably the information processing capacities of swarms over brains*."

Finally, "*intelligence as information processing*" is used as a metaphor to understand its evolution and ecology. The author's arguments about ecological (selective) pressures with respect to the evolution of intelligence and the complexity of ecosystems may be agreed or not.

In conclusion, while " … *the brain as computer metaphor is not appropriate for studying collective intelligence in general, nor swarm intelligence in particular* … " nonetheless since " … *computation can be understood as the transformation of information* (Gershenson, 2012), *"computers", broadly understood as machines that process information can be a useful metaphor* … "

Graham extends the comparison of brains and computers by introducing a further element of complexity. It is not a computer that needs to be compared to a brain; in fact, the functioning of the latter is better represented by a collection of interconnected computers (internet). He points to a relevant issue that is not solved by the proponents those that advance the "*strong*" brain-computer analogy, and this is the problem of information routing (how the information flows within the brain and the computers; how is directed from the input site to the output resolution).

2.12. Graham (12)

Daniel Graham analyzes the appropriateness of the computer–brain metaphor (see also Graham, 2021); instantiated as what he calls the "representational" view of neural components. According to Graham, the analogy is useful but incomplete. Although he agrees that the brain performs some "computations", he posits that brains themselves can be seen as the result of both representation and communication activities. The emphasis on pure mathematical operations in the brain, along with their translations in neuronal patterns of electrical spikes, does not provide a complete view of what happens inside brains as they perform tasks. One of the reasons for not supporting the strict computer–brain functional (representational) analogy is that it does not deal with the key problem of information flow within the neuronal nets of the brain. The routing of information and the remodeling of circuits transcend the limits of the computer analogy. Graham suggests the internet as a better image of our (or any) brain architecture and functional properties. The internet is constructed with clear routing protocols, with an efficient distribution of information (termed the small-network configuration, Sporns and Honey, 2006; but see also Hilgetag and Goulas, 2016 for a critical view) and the continuous remodeling of their connectivity (plus growth). These properties should remind us of the way our brains seem to be constructed and how they route information, from external/internal inputs to higher integrative circuits and on to motor systems. Integrating views of signal processing and routing strategies should give us a more nuanced view of the activities of brains.

3. Discussion

3.1. The brain-computer analogy: *"Cum grano salis"*

Much the same way as scientific hypotheses, metaphors and analogies are transitory, always adjusting to technological advances. The Brain-Computer is usually referred to as a metaphor, but it should be thought of as an analogy instead. Indeed, here we are suggesting that metaphor and Analogy are two distinct concepts and must not be used as synonyms (see also above). While we think that a salient feature of a metaphor is a "visual insight" (an evocative visual image), the concept of analogy would be mainly associated to the idea of "function." In short, metaphors have no explanatory power, whereas analogies do, for the knowledge acquired on the functionality of a system can be transferred to an analogous one, thereby leading to understanding and discovery.

The Brain-Computer analogy has raised a harsh debate in the scientific community; some took it literally whereas the very meaning of analogies implies only a partial overlap of properties. In fact, it is very possible that analogies or metaphors are inescapable (and used regularly as cognitive tools; *sensu* Lakoff and Johnson, 1980; Gomez-Marin, 2022). Metaphors are rooted in things we know and/or manipulate. In this sense the only way to grasp what many things are is by describing the phenomena in terms we understand. In this process from "the physical phenomenon" to "the understanding of it," a metaphor/analogy always arise. We hypothesize that we can claim "as original" only those things apprehended by the senses (always assuming that our senses are not tricking us). Metaphors *might be* the only things that we "comprehend," and this is because they are rooted in our sensible experiences. Kuhn himself seems to acknowledge the importance of metaphors when he claims: "*Metaphors play an essential role in establishing links between scientific language and the world. Those links are not, however, given once and for all. Theory change, in particular, is accompanied by a change in some of the relevant metaphors and in the corresponding parts of the network of similarities through which terms attach to nature*" (1993).

In this context, the wrong question to ask is if metaphors and analogies are actually useful or misleading. Quite appropriately, Yohan (2012) points out that "… *No one can claim to know how metaphors work … how we form them, and how we decide whether they are successful or not*." Along the same line, we do believe that it is totally irrelevant, to their role in science, whether metaphors and analogies are *"right or wrong*." This attitude being best exemplified by the famous Niels Bohr's horseshoe anecdote (many similar versions are available on the internet): *A friend asked if he believed in it. "Absolutely not! Bohr replied, but they say it works even if you do not believe in it*."

Usefulness seems a more appropriate adjective for metaphors; being as successful as they provide clues to the phenomena under analysis.

Moreover, from the mathematician's standpoint, but easily translatable to any scientific discipline, William Byers maintains that *"many important mathematical ideas are metaphoric in nature"* and emphasizes *"the close relationship between metaphors and ideas. A metaphor, like an idea, arises out of an act of creativity"* (Byers, 2010, p. 240). Moreover, Byers points out that "… *In general, most sweeping conjectures turn out to be "wrong" in the sense that they need to be modified during the period in which they are being worked on. Nevertheless, they may well be very valuable. The whole of mathematical research often proceeds in this way—the way of inspired mistakes. … Ideas that are "wrong" can still be valuable*."

A number of articles in this SI deal with the use of metaphors in a specific area of science, the interphase between neuroscience and computer science. In this context it is important to emphasize once more that the metaphors are essential, but also transitory, in the sense that more, newer, data to the formulation of others (or a more refined version of previous) that seem more suitable at the moment. In addition, and in the absence of new data piling up, sociological or epistemic changes could also be, at certain moments, fruitful sources of new metaphors. Moreover, Gershenson reminds us that "*different metaphors can be useful for different purposes … and in different contexts*" within a discipline. Utilizing a unique metaphor (as we explain below) might not be the most productive avenue to explain certain complex structures, for instance the brain.

In the preceding paragraphs, we have proposed two features that may be useful to characterize and differentiate metaphor from analogy. In our view, a metaphor develops from a visual image, a picture that serves as creative force for scientific insight. Again with reference to mathematics, Ivar Ekeland stresses the relationship between mathematical ideas and "certain pictures" and the power of those pictures "… *of certain visual representations, in the historical development of science … It is a power, in the early stages, to initiate progress, when the ideas it conveys are still creative and successful, and it becomes, later on, power to obstruct, when the momentum is gone and repetition of the old theories prevents the emergence of new ideas*" (Ekeland, 1988, p. 9).

As Denis Nobles put it "… *Different, even competing, metaphors can illuminate different aspects of the same situation, each of which may be correct even though the metaphors themselves may be incompatible. … Metaphors compete for insight, and for criteria like simplicity, beauty, creativity … Metaphor invention is an art not a science and, as with other art forms, the artist is not necessarily the best interpreter*" (Noble, 2006). To these views we subscribe fully.

As for analogy, we think that the criterion of "function" could be regarded as its most characterizing property, a property endowed with explanatory power. For example, in this SI, Daniel Graham proposes internet as a new metaphor for the Brain. According to our definitions (see above), in the case internet works both as metaphor and analogy.

To end this section, a cautionary note is warranted. Metaphors may be inevitable and necessary to Science, because of psychological factors associated to learning or the search for explanations (Hoffman, 1980). The alternative to using metaphors would be a crude description of facts. In the philosophy of mind (or our discussion here) this would imply a "pure" description of physiological states in the brain. Whether there is any "information content" attached to this description, the authors of this review think there is very little, if any. We do not envision a productive substitute for the use of metaphors in science.

3.2. Theories of the brain

We would like to emphasize here the importance of considering the brain as a structure that can be analyzed at "*different scales*," where functions might be distributed in particular domains and with the involvement of different components. In this sense we believe that it is wrong to search for "*an ultimate THEORY of the brain*," since a better description would have to accommodate explanations on how these levels of architectural organization (including the varied set of functional domains) are established and integrated. As explained below, it may be more appropriate to explore *"different theories of the brain*," perhaps a more suitable name for the exploratory endeavor we propose next.

In this context we would like to bring about a different perspective for analyzing the brain-computer analogy and this is through a systems approach in which the different levels of organization are candidates for specific analogies. We think that a theory that tries to analogize components (or modules, see below) should be more productive that a single theory encompassing such a complex structure as the brain. The underlying assumption here is that brains are the (non-linear) sum of components (modules) that are juxtaposed to perform, or facilitate, certain mental tasks. This is not a gratuitous assumption, since current data in the neurosciences, has proven the modularity of many of the structures in the brain, all products of evolutionary history. From the commonalities of neuronal subtypes to the conservation of specific neuronal circuits or the distribution of cortical areas, the brains of many animals share structures that were selected for specific functions and that are now recognized as homologous across taxa (Schlosser, 2018; Barsotti et al., 2021; Tosches, 2021).

In fact, brains, as any other organ or tissue, are organized at different levels, with modular blocks contributing to the next one; this suggests a parallel discussion between analogizing brain structures and the more classical discussion of biological homologies across scales (proteins, cells, organs, etc.).

In this framework, the fact that brains are organized in a series of hierarchical levels allows us to re-focus our attention on finding analogies that best represent different scales (i.e., a computer at one level, a radio at another, a hologram below, and an internet above, etc.). This does not imply setting aside the problem of the brain as a whole, just that it could be more productive finding good (useful) analogies of those, lower level, modules involved in its construction; and use modules as recognized functional units (e.g., neurons or neuronal circuits). A cautionary note needs to be introduced here: we are not claiming any strong/rigid interpretation of the brain modularity since we understand that there are clear instances of distributed function, and plasticity, plus the shifting localization of representations of stimuli over time. In fact, it is the distributed and flexible organization of modules what allows us for the integration of levels.

In a sense, and as explained by Cobb, our hypotheses on brain development and function have depended, at every historical time, on the current knowledge of the system. Hydraulic or electrical images of the brain were suggested at the time when discoveries were made on these areas; within and without the body. Calculations and algorithms promoted computational ideas of the brain, though, later on, of the neurons themselves (including the more recent idea that single neurons are doing complex computations at the synapsis). Analogies, sometimes, are exported from one level of analysis to another, with the computing image of whole brains or "single" synapsis as an obvious example. Similarly, the network analogy flows from the local connection of a few neurons (the reflex arc) to areas of the brain involved in specific tasks, to the whole brain or swarms of them.

To sum up what is explained above, we would like to suggest a reappraisal of the use of our analogies, so we can better understand how every level is organized and, importantly, how the integration of different modules at a particular level contribute to generate (emergent) properties observed at the next higher level. Moreover, we would encourage the introduction of different analogies that best represent the different levels of construction (avoiding, perhaps, the trap of overarching analogies explaining every single component in a complex system). Surely enough, we should notice here that *"parts"* in the construction of the living organism can be attributed to chance (drift), to physical–chemical laws (self-organizing), to emergent phenomena or to adaptive processes. All of these, constructional, principles bear no relationship with our purely structural view of the organism. Here we base our suggestion on the analysis of structures *per se*, at many levels, but not about their developmental assembly. Perhaps this last approximation can be considered in the future, in refined forms of our analogical search.

3.3. Creativity in science

Why are we interested in creativity in the context of this Special Issue? Because if we ask "can computers think," next we ought to ask "can computers create." And the very act of creation (be it in sciences or in the arts) stems from the awareness of the esthetic element. Reflection on creativity, and its sources, has a long history. While philosophers are far from a consensus definition of what creativity is and what it entails (see Erden, 2010), there are some tenets that are commonly recognized as pertaining to the creative act (freedom, potential, originality, etc.). Moreover, some philosophers also recognize that creation, in fact, it is a process with some specific requirements (McGilchrist, 2021): (i) a generative faculty (allowing ideas to come about; recognizing patterns, etc.) (ii) a permissive element (generating the conditions for the ideas to develop), and (iii) a translational disposition (the insights carried over for a period up to the moment of the final "creative act"). These are not, *per se*, components of other (non-necessarily creative) process such as problem solving. In the latter sense, it has been manifested that "there is no algorithm behind creative processes." What many authors agree upon is the fact that *metaphors* expand your creative thinking, and in that sense, the analysis of metaphors becomes a key component in understanding creativity in science. But how?

The analysis of metaphors and analogies relies on the understanding of its sources. Creativity is an obvious candidate. But what is the source of "*true creativity*" … in science? Undeniably, the history of science tells us that chance has played and will play an important role in scientific creativity. Aside from orthodox views, more innovative paths have been explored in recent times. This is a subject that bears an important place in our discussion about the brain-computer metaphor.

But first we need to define what creativity is and what is the suggested relationship between creativity and metaphors?

In formal and natural sciences, the issue of creativity has been thoroughly discussed mainly in Mathematics and Physics. The mathematician William Byers distinguished two types of thought in mathematics: algorithmic (based on logical operations) defined as trivial and profound (deep) thought, defined as creative (Byers, 2010).

Byers asks a number of deep questions "*Could a computer be programmed to distinguish between the trivial and the deep?*" … "*Can a computer do mathematics?*" … "*Is mathematics algorithmic?*." Therefore, and inevitably, he is confronted with "THE question" first addressed by Turing in 1950: "*Can a computer think?*," which, he says "… *is equivalent to the question: Is [human] thought algorithmic?*" If, following Byers again, "*human creativity involves ideas, ambiguity, paradox, depth, and complexity*" an act of creativity (a very rare event indeed) might be analogous to a biological evolutionary event, since (as some authors have pointed out), it is impossible to predict, *a priori*, how ideas will unfold in the future. Ideas seem to unfold over time, through culture (Gabora and Kaufman, 2010).

The mathematician Henri Poincaré, in asking "*what is mathematical creation?*" proposed the following: "*The mathematical facts worthy of being studied are those which, by their analogy with other facts, are capable of leading us to the knowledge … Among chosen combinations the most fertile will often be those formed of elements drawn from domains which are far apart … Invention is discernment, choice*" (Poincaré, 1910). The theoretical physicist Paul Dirac, followed an even more unconventional path designating the potent role of esthetics in scientific creativity (see below). We share the opinion of those authors who think metaphors to be invaluable to scientific creativity in that they permit to explore uncharted lands and offer new perspectives for expanding scientific knowledge.

Here we propose to consider another aspect of creativity that we think to be highly relevant, but still not fully appreciated: it is the feeling of emotion, of amazement (*émerveillement*, in French) the researcher may feel in front of a phenomenon (though Socrates and Plato linked it already to wisdom). This feeling might be necessary and sufficient to awaken the scientist and illuminate his/her thoughts. Moreover, and probably facilitated by "*émerveillement*," we would also like to suggest that creativity in science might depend on the integration of views arising from different disciplines (as many scientists have stated when asked about their own work). In this context, we would speculate that "transferring metaphors" from one field to another could be a good source of new ideas, thus propitiatory of a creative act. Metaphors as tools for understanding in one field should be able to illuminate other aspects of reality, in another field. This mental "transferring" could be productive, and, thus lead to "understanding" of unrelated phenomena.

In the context of the Brain-Computer analogy the comparison between human creativity and "AI creativity" arises spontaneously. And, the debate as to whether AI does and will play a role in human creativity is undeniably timely.

Obviously, human creativity is influenced by cultural traditions (context), or through the connections between very different ideas (i.e., from different/distant fields). This might suggest some intrinsic difficulties in imitating (the process of) human creativity with AI.

In a rather trivial way, and following the assertion above, we can state that computers are not creative, unable to produce "acts of creativity," since creation might not be a pure algorithmic process.

Indeed, supporting the preeminence of human creativity, Byers (2010) writes "… *mathematical thought can be simple and it can also be complex but mostly it is nontrivial. Computer thought, on the other hand, even though it may be very lengthy and complicated, is essentially trivial.*" Along the same lines, Roli et al. support the notion that creativity in mathematics "… *does not come out of algorithmic thought but via insight,*" based on the argument that AI algorithms (being human devised) cannot identify, appraise and exploit (adapt to) new "environmental" alternatives (called *affordances*), to new frames of reference (Roli et al., 2022*).

Advocates of the potential of AI describe its performance, either in autonomous creative processes or with human intervention, mainly in fields such as writing, music, and painting (e.g., Zylinska, 2020; Jukebox, https://openai.com/blog/jukebox/), but also in scientific discovery (e.g., protein fold prediction *via* machine-and deep-learning techniques; see Gil et al., 2014 and Callaway, 2020).

Other authors sustain a more optimistic view of AI creativity (e.g., Boden, 2003; Forbes, 2020). These authors, following tenets of classical psychology (e.g., Boden, 1992), view creativity as having different sources and, thus, classify it as: combinatorial (combination of familiar ideas; e.g., poetic images), exploratory (arising within cultural traditions; e.g., cooking recipes), and transformational creativity (older rules are broken; e.g., cubism). While combinatorial and exploratory components are imitable by AI, the possibility of transformational creativity in computing seems a bit more problematic. Breaking rules seem to be a capacity specific to humans, since in the final acceptance of those new rules depends on a "value judgement" (only those outcomes that satisfy some new needs are incorporated). Of course, the possibility of having evolving programs (with selective regimes) are now a reality and this allows the possibility of "transformation." Here again value is an important selective factor, which can hardly be implemented (nowadays) by AI. We should be reminded that critical thinking is still central to the creative process and, while humans are able to scrutinize their ideas/creations, computers cannot. All in all, these authors seem to suggest the possibility of finally exorcizing Cartesian dualism while establishing that the brain is, perhaps, a wonderfully subtle machine.

In spite of these, sometimes, entrenched opinions, what every scientist seems to agree upon is that with the current state of AI, our "machines" capabilities for imitating human creativity is still quite limited nowadays, though the future can change our views quite rapidly. All in all, creativity should be incorporated as a key concept when discussing metaphors such as that relevant to this Special Issue.

To end this review essay, we think it is important to mention the role of beauty (and esthetics in general) in scientific creativity.

Creativity and esthetics go hand in hand. Evaluation of creativity always requires a judgment of beauty. For instance, and according to a Kant's classical statement: "aesthetics is not the goal of creativity but it is its essential component."

Its crucial role in science has been recognized by a number of great mathematicians and physicists: G. W. Leibniz, H. Poincaré, A. Einstein, G. H. Hardy, P. A. M. Dirac, M. Gell-Mann, to name just a few. In 1910 Poincare writes "*… the feeling of mathematical beauty, of the harmony of numbers and forms, of geometric elegance. This is a true esthetic feeling that all real mathematicians know, and surely it belongs to emotional sensibility. Now, what are the mathematic entities to which we attribute this character of beauty and elegance, and which are capable of developing in us a sort of esthetic emotion?*" (Poincaré, 1910). In a famous and often quoted sentence Dirac boldly stated that "*… it is more important to have beauty in one's equations than to have them fit experiment*" (Dirac, 1963). And in G. H. Hardy's words: "*The mathematician's patterns, like the painter's or the poet's, must be beautiful.*" Finally, the pervasiveness of the esthetics approach to human knowledge is epitomized in two extraordinarily powerful lines by the romantic English poet John Keats: "*What the imagination seizes as beauty must be the truth*" (1817) and "*Beauty is truth, truth beauty. That is all ye know, and all ye need to know*" (1884; in Keats, 2015). The relationships linking computer science, creativity and esthetics are explored in a recent review by Yang and Lu (2022), in which the authors also propose a framework that uses computational methods to connect creativity and esthetics.

It is fairly obvious that the meaning of the concept of beauty differs greatly among cultures, and also among individuals; this is trivial. Indeed, the commonly accepted stance is that the notion of esthetic quality is elusive. But here lies the problem, as metaphorically illustrated by the famous Bruegel's painting "*The blind conducts the blind*"; everybody is looking for the "absolute" definition of beauty, and exactly this is the mistake. In contrast, the most crucial point is that, as far as science is concerned, all meanings and definitions of beauty are equivalent, and they all fulfill the same goal: to show the way to scientific discovery. Beauty is impossible to define because it lies in the eyes of the observer? Sure, but … "*It does not matter!*" for the chances for any esthetic criterion to be effective are not negligible at all. Likewise, for metaphors and analogies.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Acknowledgments

We thank Matthew Cobb and Alex Gomez-Marin for critical reading of the MS. We also thank R. Brette, M. Cobb, M. Chirimuuta, A. Danchin, C. Gershenson, A. Gomez-Marin, D. Graham, M. Levin, B. Richards, A. Roli, and R. Solé for feedback on resumes of their articles. Thanks are also due to Marino Miculan for clarifying a number of concepts in computer science. GM wishes to thank most warmly Annalisa Pastore and Pierandrea Temussi for valuable advice and constant encouragement.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Atran, S. (1990). *Cognitive Foundations of Natural History, Towards an Anthropology of Science.* Cambridge, UK: Cambridge University Press.

Barsotti, E., Correia, A., and Cardona, A. (2021). Neural architectures in the light of comparative connectomics. *Curr. Opin. Neurobiol.* 71, 139–149. doi: 10.1016/j.conb.2021.10.006

Black, M. (1962). *Models and Metaphors, Studies in Language and Philosophy.* Ithaca, New York, USA: Cornell University Press.

Boden, M. A. (1992). Understanding creativity. *J. Creat. Behav.* 26, 213–217. doi: 10.1002/j.2162-6057.1992.tb01178.x

Boden, M. A. (2003). *The Creative Mind: Myths and Mechanisms.* Milton Park, England: Routledge.

Bongard, J., and Levin, M. (2021). Living things are not (20th century) machines: updating mechanism metaphors in light of the modern science of machine behavior. *Front. Ecol. Evol.* 9:650726. doi: 10.3389/fevo.2021.650726

Boullerne, A. I., Foley, P., Turner, A. J., Johnston, G. A. R., and Beart, P. M. (2020). The origins and early history of neurochemistry and its societies. *J. Neurochem.* 152, 8–28. doi: 10.1111/jnc.14839

Brady, S., Siegel, G., Albers, R. W., and Editors P. D. L. (2011). *Basic Neurochemistry: Principles of Molecular, Cellular and Medical Neurobiology.* New York, USA: Academic Press.

Brette, R. (2022). Brains as computers: metaphor, analogy, theory or fact? *Front. Ecol. Evol.* 10:878729. doi: 10.3389/fevo.2022.878729

Byers, W. (2010). *How Mathematicians Think*. Princeton, USA: Princeton University Press.

Callaway, E. (2020). It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* 588, 203–204. doi: 10.1038/d41586-020-03348-4

Chirimuuta, M. (2022). Artifacts and levels of abstraction. *Front. Ecol. Evol.* 10:952992. doi: 10.3389/fevo.2022.952992

Cobb, M. (2020). *The Idea of the Brain: A History*. New York, NY, USA: Basic Books.

Cobb, M. (2021). A brief history of wires in the brain. *Front. Ecol. Evol.* 9:760269. doi: 10.3389/fevo.2021.760269

Cole, D. (2020). "The Chinese room argument," in *The Stanford Encyclopedia of Philosophy (Winter 2020 Edition)*. ed. E. N. Zalta. (Stanford, USA)

Danchin, A., and Fenton, A. A. (2022). From analog to digital computing: is Homo sapiens' brain on its way to become a Turing machine? *Front. Ecol. Evol.* 10:796413. doi: 10.3389/fevo.2022.796413

Darwin, C. (1859). *The Origin of Species By Means of Natural Selection*. London: Murray.

Davidson, E. H., and Peter, I. (2015). *Genomic Control Process: Development and Evolution.* New York, USA: Academic Press.

Davis, M. (2021). The brain-as-computer metaphor. *Front. Comp. Sci.* 3:681416. doi: 10.3389/fcomp.2021.681416

Dirac, P. A. M. (1963). The evolution of the physicists picture of nature. *Sci. Am.* 208, 45–53. doi: 10.1038/scientificamerican0563-45

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science* 284, 2124–2128. doi: 10.1126/science.284.5423.2124

Ekeland, Y. (1988). *Mathematics and the Unexpected.* Chicago, USA: University of Chicago Press.

Epstein, R. (2016). Your brain does not process information and it is not a computer, Aeon. Available at: https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer

Erden, Y. J. (2010). Could a created being ever be creative? Some philosophical remarks on creativity and AI development. *Mind. Mach.* 20, 349–362. doi: 10.1007/s11023-010-9202-2

Faisal, A. A., Selen, L. P., and Wolpert, D. M. (2008). Noise in the nervous system. *Nat. Rev. Neurosci.* 9, 292–303. doi: 10.1038/nrn2258

Fernandez, N., Maldonado, C., and Gershenson, C. (2014). "Information measures of complexity, emergence, self-organization, homeostasis, and autopoiesis," in *Guided Self-Organization: Inception. Vol. 9.* ed. M. Prokopenko (Berlin Heidelberg: springer).

Forbes, A. C. (2020). AI: from expressive mimicry to critical inquiry. *Artnodes* 26, 1–10. doi: 10.7238/a.v0i26.3370

Fraser, P., Solé, R., and De las Cuevas, G. (2021). Why can the brain (and not a computer) make sense of the liar paradox? *Front. Ecol. Evol.* 9:802300. doi: 10.3389/fevo.2021.802300

Gabora, L., and Kaufman, S. B. (2010). "Evolutionary perspectives on creativity," in *The Cambridge Handbook of Creativity*. eds. J. Kaufman and R. Sternberg (Cambridge, UK: Cambridge University Press), 279–300.

Gentner, D., and Jezioreski, M. (1993). "The shift from metaphor to analogy in Western science," in *Metaphor and Thought. 2nd Edn*. ed. A. Ortony (Cambridge, UK: Cambridge University Press), 447–480.

Gershenson, C. (2012). "The world as evolving information," in *Unifying Themes in Complex Systems. Vol. VII.* eds. A. Minai, D. Braha and Y. Bar-Yam (Berlin Heidelberg: Springer), 100–115.

Gershenson, C. (2021). Intelligence as information processing: brains, swarms, and computers. *Front. Ecol. Evol.* 9:755981. doi: 10.3389/fevo.2021.755981

Gil, Y., Greaves, M., Hendler, J., and Hirsh, H. (2014). Amplify scientific discovery with artificial intelligence. *Science* 346, 171–172. doi: 10.1126/science.1259439

Gomez-Marin, A. (2022). Commentary: metaphors we live by. *Front. Comp. Sci.* 4:890531. doi: 10.3389/fcomp.2022.890531

Graham, D. (2021). *An Internet in Your Head: A New Paradigm for How the Brain Works*. New York: Columbia University Press.

Haack, S. (2019). The art of scientific metaphors. *Rev. Port. Filos.* 75, 2049–2066. doi: 10.17990/RPF/2019_75_4_2049

Hatfield, G. (2012). "Mechanizing the sensitive soul," in *Matter and Form in Early Modern Science and Philosophy*. ed. G. Manning (Leiden, Netherlands: Brill).

Hawkins, J., and Blakeslee, S. (2004). *Times Books*. Chicago, USA: Times Books.

Hesse, M. B. (1970). *Models and Analogies in Science.* Notre Dame, USA: University of Norte Dame Press.

Hilgetag, C. C., and Goulas, A. (2016). Is the brain really a small-world network? *Brain Struct. Funct.* 221, 2361–2366. doi: 10.1007/s00429-015-1035-6

Hoffman, R. R. (1980). "Metaphor in science," in *Cognition and Figurative Language.* eds. R. P. Honeck and R. R. Hoffman (New York: Routledge).

Hoffman, R. R. (1985). "Some implications of metaphor for philosophy and psychology of science," in *The Ubiquity of Metaphor: Metaphor in Language and Thought. Vol. 29.* eds. W. Paprotté and R. Dirven (Amsterdam, Netherlands: Current Issues in Linguistic Theory; John Benjamins Publishing Company), 327–380.

Humar, M. (2021). Metaphors as models: towards a typology of metaphor in ancient science. *HPLS* 43:101. doi: 10.1007/s40656-021-00450-2.Jukebox

Keats, J. (2015). *The Odes of John Keats*. Australia: Leopold Classic Library.

Kerskens, C. M., and Lopez-Perez, D. (2022). Experimental indications of non-classical brain functions. *J. Phys. Commun.* 6:105001. doi: 10.1088/2399-6528/ac94be

Kovac, J. (2003). Writing as thinking. *Ann. N. Y. Acad. Sci.* 988, 233–238. doi: 10.1111/j.1749-6632.2003.tb06103.x

Kuhn, T. S. (1993). "Metaphor in science," in *Metaphor and Thought. 2nd Edn.* ed. A. Ortony (Cambridge, UK: Cambridge University Press), 533–542.

Lakoff, G. (1993a). "The contemporary theory of metaphor," in *Metaphor and Thought. 2nd Edn*. ed. A. Ortony (Cambridge, UK: Cambridge University Press).

Lakoff, G. (1993b). "The syntax of metaphorical semantic roles," in *Semantics and the Lexicon. Studies in Linguistics and Philosophy. Vol. 49.* ed. J. Pustejovsky (Dordrecht: Springer).

Lakoff, G., and Johnson, M. (1980). *Metaphors We Live By.* Chicago, USA: Chicago University Press.

Lamarck, J.-B. (1809). *Philosophie Zoologique*. *Vol. 2* (Paris, France: Dentu), 463.

Lima, M. (2014). *The Book of Trees.* New York, USA: Princeton Architectural Press.

Mahoney, M. S. (1988). The history of computing in the history of technology. *Ann. Hist. Comput.* 10, 113–125. doi: 10.1109/MAHC

Martin, W. (1999). Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *Bio Essays* 21, 99–104. doi: 10.1002/(SICI)1521-1878(199902)21:2<99::AID-BIES3>3.0.CO;2-B

Martinez, P., and Sprecher, S. G. (2020). Of circuits and brains: the origin and diversification of neural architectures. *Front. Ecol. Evol.* 8:82. doi: 10.3389/fevo.2020.00082

Maxwell, J. C. (1870). "Address to the mathematical and physical sections of the British association. British association report, p 227; reprinted," in *The Scientific Papers of James Clerk Maxwell*. ed. W. D. Niven (Cambridge, UK: Cambridge University Press), 215–229.

McCulloch, W., and Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133. doi: 10.1007/BF02478259

McGilchrist, I. (2021). *The Matter With Things: Our Brains, Our Delusions, and the Unmaking of the World*. UK: Perspectiva Press.

Nicholson, D. J. (2013). Organisms ≠ Machines. *Stud. Hist. Philos. Biol. Biomed. Sci.* 44, 669–678. doi: 10.1016/j.shpsc.2013.05.014

Noble, D. (2006). *The Music Of Life-Biology Beyond Genes.* Oxford, UK: Oxford University Press.

Owen, R. (1843). *Lectures on the Comparative Anatomy and Physiology of the Invertebrate Animals*. London: Longman, Brown, Green, and Longmans.

Pitts, W., and McCulloch, W. S. (1947). How we know universals the perception of auditory and visual forms. *Bull. Math. Biophys.* 9, 127–147. doi: 10.1007/BF02478291

Poincaré, H. (1910). Mathematical creation. *Monist* 20, 321–335.

Ragan, M. A. (2009). Trees and networks before and after Darwin. *Philos. Trans. R. Soc. B* 364, 2169–2175. doi: 10.1098/rstb.2009.0046

Reynolds, A. S. (2022). *Understanding Metaphors in the Life Sciences (Understanding Life).* Cambridge, UK: Cambridge University Press.

Richards, B. A., and Lillicrap, T. P. (2022). The brain-computer metaphor debate is useless: a matter of semantics. *Front. Comp. Sci.* 4:810358. doi: 10.3389/fcomp.2022.810358

Roli, A., Jaeger, J., and Kauffman, S. A. (2022). How organisms come to know the world: fundamental limits on artificial general intelligence. *Front. Ecol. Evol.* 9:806283. doi: 10.3389/fevo.2021.806283

Schlosser, G. (2018). A short history of nearly every sense-the evolutionary history of vertebrate sensory cell types. *Integr. Comp. Biol.* 58, 301–316. doi: 10.1093/icb/icy024

Searle, J. (1984). *Minds, Brains and Science.* Cambridge MA: Harvard University Press.

Sporns, O., and Honey, C. J. (2006). Small worlds inside big brains. *Proc. Natl. Acad. Sci. U. S. A.* 103, 19219–19220. doi: 10.1073/pnas.0609523103

Sterling, P., and Laughlin, S. (2017). *Principles of Neural Design*. Cambridge, MA: MIT Press.

Tosches, M. A. (2021). Different origins for similar brain circuits. *Science* 371, 676–677. doi: 10.1126/science.abf9551

Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proc. Lond. Math. Soc.* s2-42, 230–265. doi: 10.1112/plms/s2-42.1.230

Turing, A. M. (1950). Computing machinery and intelligence. *Mind* LIX, 433–460. doi: 10.1093/mind/LIX.236.433

von Neumann, J. (1958). *The computer and the Brain.* New Haven, USA: John Yale University Press.

Waddington, C. H. (1957). *The Strategy of the Genes; A Discussion of Some Aspects of Theoretical Biology.* Crows Nest, Australia: George Allen & Unwin Ltd.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* 16, 97–159. doi: 10.1093/genetics/16.2.97

Yang, H., and Lu, Z. (2022). "Computerising connections between creativity and aesthetics," in *IEEE International Conference on Service-Oriented System Engineering (SOSE)*, 185–188.

Yohan, J. J. (2012). Metaphor: the Alchemy of Thought. Available at: https://axispraxis.wordpress.com/2012/09/26/metaphor-the-alchemy-of-thought/

Ziman, J. (2000). *Real Science, What It Is and What It Means.* Cambridge, UK: Cambridge University Press.

Zylinska, J. (2020). *AI Art-Machine Visions and Warped Dreams*. London: Open Humanities Press.

PERSPECTIVE
published: 16 December 2021
doi: 10.3389/fevo.2021.760269

A Brief History of Wires in the Brain

*Matthew Cobb**

School of Biological Sciences, The University of Manchester, Manchester, United Kingdom

Metaphors have formed a significant part of the development of neuroscience, often linked with technology. A metaphor that has been widely used for the past two centuries is that of the nervous system being like wires, either as a telegraph system or telephone exchange, or, more recently, in the more abstract metaphor of a wiring diagram. The entry of these terms into scientific writing is traced, together with the insights provided by these metaphors, in particular in relation to recent developments in the study of connectomes. Finally, the place of the wiring diagram as a modern version of Leibniz's "mill" argument is described, as a way of exploring the limits of what insight the metaphor can provide

Keywords: wiring diagram, brain, neuroscience, history, metaphor

OPEN ACCESS

Edited by:
Giorgio Matassi,
FRE3498 Ecologie et dynamique des systèmes anthropisés (EDYSAN), France

Reviewed by:
Volker Hartenstein,
University of California, Los Angeles, United States
Anil K. Seth,
University of Sussex, United Kingdom

****Correspondence:***
Matthew Cobb
cobb@manchester.ac.uk

Specialty section:
This article was submitted to Models in Ecology and Evolution, a section of the journal Frontiers in Ecology and Evolution

Received: *17 August 2021*
Accepted: *26 November 2021*
Published: *16 December 2021*

Citation:
Cobb M (2021) A Brief History of Wires in the Brain. Front. Ecol. Evol. 9:760269. doi: 10.3389/fevo.2021.760269

INTRODUCTION

Our changing understanding of brain function has involved the use of metaphors, often taken from technology (Cobb, 2020). The role of metaphors in science has been studied by philosophers (e.g., Lakoff and Johnson, 1980; Brown, 2003); metaphors shed light on phenomena but also frame and sometimes limit how we can think about them. In this Perspective I explore the metaphor of "wiring" in the brain, the insights it provides and the scientific and conceptual issues raised by this metaphor, some of which go back to 18th century debates and are still unresolved today.

FROM HYDRAULICS TO THE TELEGRAPH

In the Western tradition, it was thought for millennia that movement was produced by a fluid or spirit in the nerves, coming from the heart or, according to some minority views, the brain. By the 1630s, when it was understood that the heart was merely a pump while the brain was anatomically highly complex, Descartes suggested that movement and brain function occurred through a hydraulic mechanism, similar to that he observed in moving statues in Parisian parks. But sectioning nerves showed there was no such fluid. This left thinkers at a loss; in the 1670s the pioneer microscopist Jan Swammerdam suggested that whatever moved down a nerve might be like a vibration travelling down a plank of wood, but he could not suggest how this might work (Swammerdam, 1758). At the time, most ideas about brain function used mechanical metaphors – the term "impression", still in everyday use, implied that stimuli pushed upon the structures of the brain, leaving their shape – an impression. Despite their power and longevity, these ideas failed the basic test of science – there was no evidence for them.

The mastery of electricity in the second half of the 18th century allowed precise experimentation on both isolated nerves and eventually on the brain, leading to new, more informative metaphors regarding brain function. It also had a contradictory effect – because the language of electricity is based on watery metaphors (current, flow, etc.), aspects of our thinking of brain function are pulled back to the old hydraulic metaphors. More significantly, with the development of the telegraph system in the late 1830s a powerful parallel was drawn: the nervous system was described as being like a telegraph, while the telegraph system was seen as the nervous system of the country. Both telegraph and nerves involved near-instantaneous communication and they both enabled action.

For the mid-century inventor Alfred Smee, the nervous system was literally telegraphic: "In animal bodies we really have electro-telegraphic communication in the nervous system. That which is seen, or felt, or heard is telegraphed to the brain" (Smee, 1850). Many thinkers suggested that the same kind of stuff was going down both wires and nerves – "intelligence", or as Dr. Spencer Thomson put it: "the wires – nerves – convey the information from all parts of the body" (Thomson and Smith, 1853).

A few years later, in 1863, the great German physiologist Hermann von Helmholtz pointed out that nerves, like telegraph wires, could produce all sorts of functions: "Nerves have often and not unsuitably been compared to telegraph wires. according to the different kinds of apparatus with which we provide its terminations, we can send telegraphic dispatches, ring bells, explode mines, decompose water, move magnets, magnetise iron, develop light, and so on. So with the nerves." (Helmholtz, 1875). Helmholtz argued that the differences in the activity of different parts of the nervous system (for example, different sensory modalities), were not due to what his teacher, Johannes Müller, had called "the law of specific nerve energies". Helmholtz argued that all nerves carried the same kind of signal, and that different sensations arose when the brain interpreted them in different ways.

Thirty years later, Ramon y Cajal used the telegraph network to explain the structure and function of a single neuron: "The nerve cell consists of an apparatus for the *reception* of currents, as seen in the dendritic expansions and the cell body, an apparatus for *transmission*, represented by the prolonged axis cylinder, and an apparatus for *division* or *distribution*, represented by the nerve terminal arborisation." (Cajal, 1894). Cajal even used wiring as a way of explaining what was happening in the as yet unnamed synapse: "current must be transmitted from one cell to another by way of *contiguity* or *contact*, as in the splicing of two telegraph wires" (Robinson, 2001).

Nevertheless, Cajal felt that the telegraph was not a precise model for how the brain worked. Brains were plastic, unlike the fixed telegraph: "A continuous pre-established network – a kind of grid composed of telegraph wires in which neither new nodes nor new lines can be created – is something rigid, immutable, incapable of being changed, which clashes with the widespread impression that the organ of thought is, within certain limits, malleable and capable of perfection, above all during its development, by means of well-directed mental exercise." (Cajal, 1894).

FROM SWITCHBOARDS TO WIRING DIAGRAMS

Toward the end of the 19th century a new technology challenged the rigid telegraph metaphor – the telephone exchange, where messages can be flexibly routed. For French philosopher Henri Bergson "the brain is no more than a kind of central telephonic exchange: its office is to allow communication, or to delay it . . . it really constitutes a centre, where the peripheral excitation gets into relation with this or that motor mechanism" (Bergson, 1911). The telephone exchange metaphor is still occasionally used in popular writing – for example, in 2014, Stanislas Dehaene wrote: "consciousness is nothing but the flexible circulation of information within a dense switchboard of cortical neurons" (Dehaene, 2014). However, the limitations of the flexibility seen in a switchboard – they do not even contain simple feedback loops – mean that richer metaphors have often been favoured over the last half century.

In the final years of the 19th century, planning and recording the cabling of a telephone exchange, a telegraph system or even a house led to the appearance of the term "wiring diagram". In 1912 the British surgeon Deane Butcher used the metaphor of wiring in a house to describe the innervation of a muscle cell (Butcher, 1912), while one of the first applications of the term "wiring diagram" to the nervous system came in 1922, by Harvard psychologist Leonard Troland: "From the retina to the brain and hence from the retina to the visual consciousness, the process of seeing depends upon an extremely intricate telegraphic system. It is essential that we should determine the "wiring diagram" of this system for human beings and for any animal species, the visual processes of which we may be studying." (Troland, 1922).

The idea that form can cast light on function gained impetus in the 1940s, following the influential but mistaken suggestion by McCulloch and Pitts (1943) that different forms of synapse expressed what they called "the immanent logic of the nervous system". With the development of computers following the work of von Neumann, itself inspired by the logical concepts outlined by McCulloch and Pitts, the two kinds of wiring diagram – that of the animal body and that of the computer – entered into dialogue. McCulloch explained his approach: "regarding the anatomy of the nervous system as if it were a wiring diagram and the physiology of the neuron as if it were a component relay of a computing machine, we shall describe the brain in terms thoroughly familiar to the electrical engineer whose business is communication." (McCulloch and Pfeiffer, 1949).

In the 1950s, the development of early computer models of pattern recognition reinforced the idea of a parallel between wiring in machines and humans. In 1958, psychologist Frank Rosenblatt argued that "if one understood the code or "wiring diagram" of the nervous system, one should, in principle, be able to discover exactly what an organism remembers by reconstructing the original sensory patterns from the "memory traces" which they have left" (Rosenblatt, 1958). Nevertheless, there were clear limits to the precision of the wiring diagram metaphor, because of the plasticity and distributed function of most nervous systems. As Pitts pointed out: "it is never predetermined that a particular cell in a particular place shall project to another particular cell in another particular place, but only that all cells of a given type in a particular locality shall connect (roughly) to cells in another definite locality" (Pitts, 1955).

A more precise version of "wiring diagram" appeared with the advent of valve-based electrical circuits in the early decades of the 20th century. "Circuit diagrams", which represent not only wires but also nodes corresponding to precise functions (resistors, diodes, and so on) were soon applied with success to electro-chemical models of neuronal membrane function (e.g., Cole and Curtis, 1939; Stadler, 2017). Circuit diagrams, with their greater detail and implicit focus on function, rather than simply on routing, suggest that the overall function of the circuit may be understandable from structure.

However, until recently there were few worked examples of such interpretations of biological circuit function due to lack of anatomical and biochemical knowledge. For example, Bullock and Horridge (1965), a monumental survey of invertebrate nervous systems, contained few circuit diagrams, which rarely went beyond specifying activation or inhibition at a particular node (this had been a feature of diagrams of the brain since the middle of the 19th century). Bullock and Horridge explained that "little is actually known about specific neuronal connections", but that what was known "understandably encourages the speculation that specified circuits of some complexity are a major principle of neural function. For the most part, this is still a theoretical area".

To overcome these problems, some researchers focused on extremely simple animal systems in which the organisation and activity of single cells could be precisely described. By 1970, Eric Kandel was using "wiring diagram" to describe his work on the gill withdrawal reflex in *Aplysia*, both metaphorically and literally (Kandel, 1970). Kandel was able not only to trace the neuronal connections between the various parts of his favourite mollusc, he used circuit diagrams to explain the functional relations between the components. In the hands of Kandel and others, the circuit diagram became simultaneously a metaphor, a description and a hypothesis.

As Cajal and Pitts realised, part of the reason why a wiring diagram – or a circuit – is not an entirely accurate description of the nervous system is that the wires in your house or your computer are fixed with precise connections (or they should be), whereas in its detail the nervous system is imprecise and plastic. Over the last few decades, some researchers have used cloud computing or the internet as metaphors for brain function, with neurons or groups of neurons forming distinct functional subunits carrying out particular computations within a distributed structure (e.g., Cazé et al., 2013). The advantage of this relatively rare metaphor is that it embodies plasticity and distributed function into our conception of the brain, but explaining how exactly that distributed function works in any given case remains a challenge. Furthermore, the limits of this metaphor are quite evident: the internet is designed to function even if key parts it removed, whereas some aspects of brain function can be decisively disrupted if particular areas are damaged.

CONNECTOMES

The wiring diagram metaphor became particularly widespread with the development of various connectomic projects in the 21st century, even though the article that kicked off the interest in mammalian connectomes (Crick and Jones, 1993) did not refer to wiring diagrams at all. (The term connectome was coined separately by two researchers in 2005 – Hagmann, 2005; Sporns et al., 2005). Connectomic projects, which are aimed at a complete description of the structural connectivity of the central nervous system, can involve very different levels of resolution, depending on whether they focus on neurons or nervous tracts. For example, in 2009 the United States Human Connectome Project, which uses brain scans to describe bundles of nerves that connect brain regions, was claimed to represent "the wiring diagram of the entire, living human brain" (Bardin, 2012). But this map of macroconnections is a distinctly different kind of wiring diagram from the first connectome to be established, the 1986 cell-level description of Caenorhabditis elegans (White et al., 1986) – half-jokingly described by Sydney Brenner's laboratory as "the mind of a worm" (White, 2013). Notably, White et al. (1986) did not use the wiring metaphor once in the 340 pages of their article, preferring "circuitry".

For the moment, there is no sign of either the wiring diagram or circuit diagram metaphor going out of fashion. Even scientists who are critical of the emphasis on connectomics happily use the wiring diagram metaphor (e.g., Barack and Krakauer, 2021; Gomez-Marin, 2021). For the moment there is little reason why these metaphors should be dropped – they serve a useful function for both scientists and the general public, explaining anything from a connectome to a neural network in a simple way and suggesting a link between structure and function. As MIT neuroanatomist Lennart Heimer wrote in 1971: "In order to arrive at a detailed understanding of how the brain works we need a clear knowledge of this wiring diagram. Obviously the diagram itself could not explain the workings of the human mind, but a meaningful picture of the wiring system is a prerequisite for such understanding" (Heimer, 1971).

The limits to the metaphor are those of all biological metaphors – they are not exact descriptions of the phenomena in question. But as long as those who use them realise that there is an inevitable inexactitude at the heart of the image, no harm will be done. For the moment there is no sign of scientists being trapped by the confines of the wiring diagram metaphor, of missing potential insights because of their commitment to the metaphor, probably because "wiring diagram" is intrinsically loose and is recognised as a metaphor, if only for the obvious reason that every biologist knows that neurons are not wires. Future metaphors are hard to predict, because experience suggests they will be based on currently unknown technology. As a word of warning for those with an appetite for new metaphors, use of novel technologies as metaphors has not necessarily led to insight or to broad takeup (e.g., the suggestion that memory functions fractally, like a hologram – Pribram, 1969).

DISCUSSION

There is a further problem lurking within all representations of the brain, be they metaphorical or literal. In 1974, the psychologist Stuart Sutherland argued that even if we had "a complete wiring diagram of an individual human brain including a specification of the exact probabilities of synaptic transmissions occurring at all synapses and everything else necessary to build an exact simulation of the system. (...) it could not be claimed that we had succeeded in understanding how the brain worked; we would merely have succeeded in simulating its workings." (Sutherland, 1974). In 2012 NIH chief Francis Collins complained about the static representations produced by connectome studies: "It'd be like, you know, taking your laptop and prying the top off and staring at the parts inside, you'd be able to say, yeah, this is connected to that, but you wouldn't know how it worked." (Bardin, 2012).

This concern goes back to 1712, when the philosopher Gottfried Leibniz argued that a detailed description of the brain would not explain thought and perception, just as seeing the components of a machine does not explain how it works. This argument, which became known as Leibniz's Mill, has troubled thinkers and scientists down the ages. A wiring diagram on its own will not explain perception or virtually any other part of behaviour – individual differences in synaptic strength and organisation, which do not form part of a simple wiring diagram, can produce individual differences in behaviour (Stern et al., 2017).

In 1946, Yale physicist Roland Meyerott used the wiring diagram metaphor to address this fundamental problem – the link between structure and function:

"Many details of the functioning of the neural units are known, but how, for example, the neural units are combined in the visual area to enable the organism to locate an object seen and to act accordingly is not explained by these observational techniques. It is not likely that a "wiring diagram" of the nervous system of an organism, even if it could be uniquely traced, can ever yield this type of information. Since this is a problem involving space and time intervals, a theory based on the properties of the neural elements will be required in conjunction with a "wiring diagram" in order to explain the behaviour of the organism" (Meyerott, 1946).

The difficulty with using even a highly detailed wiring diagram to accurately predict function can be seen from studies of simple nervous systems. *C. elegans* worms at the same developmental state produce different changes in the activity of their synapses in response to starvation, leading to different responses (Bhattacharya et al., 2019). This may be partly due to subtle differences in individual connectomes that emerge during development (Witvliet et al., 2021). The circuit composed of 30-odd neurons that is found in the lobster's stomach can produce radically different behaviours, while the same behaviour can be produced by very different circuits (Bargmann and Marder, 2013). At the other end of the scale of brain complexity, human brains show the same physical connectivity but different functional configurations under anaesthesia and when awake (Barttfeld et al., 2015).

In other words, wiring diagrams, no matter how complex, are not enough. The issue highlighted by Meyerott remains: we need a theory – or theories – to explain how neural networks function. Our current theoretical approaches have been shaped by the metaphor that has dominated our thinking about the brain since the 1950s, which is that the brain is something like a computer, carrying out computations that enable it to model the present and predict the consequences of actions on future states (Cobb, 2020). This metaphor clearly involves the "wiring diagram" metaphor described here, with all its power and limits. But a distinction between the two metaphors is beginning to emerge – although "wiring diagram" retains its influence, over the last decade or so some neuroscientists have become increasingly uneasy with this starting point, expressing frustration at the wave of anatomical, genetic and electrophysiological data we are collecting without a theoretical framework (e.g., Sporns, 2015; Churchland and Abbott, 2016; Frégnac, 2017). Starting at the "top", trying to develop a theory to explain the functioning of the wiring diagram of the human or mammalian brain, seems to me to be an error. Instead, we should attempt to develop such a theory by studying small networks where we can know the precise structural, functional and effective connectivity at a cellular level and study their function using theoretical models (e.g., Friston et al., 2013). Applying such a theory to the wiring diagram of ourselves will be an immense challenge – the work of centuries, I expect. Our current inability to understand the function of the wiring in the lobster's stomach – or in the worm, or in the maggot's brain – is a measure of the task before us.

AUTHOR CONTRIBUTIONS

MC conceived and wrote the article.

REFERENCES

Barack, D. L., and Krakauer, J. W. (2021). Two views on the cognitive brain. *Nat. Rev. Neurosci.* 22, 359–371. doi: 10.1038/s41583-021-00448-6

Bardin, J. (2012). Making connections. *Nature* 483, 394–396. doi: 10.1038/483394a

Bargmann, C., and Marder, E. (2013). From connectome to brain function. *Nat. Methods* 10, 483–490. doi: 10.1038/nmeth.2451

Barttfeld, P., Uhrig, L., Sitt, J. D., Sigman, M., Jarraya, B., and Dehaene, S. (2015). Signature of consciousness in the dynamics of resting-state brain activity. *Proc. Natl. Acad. Sci.* 112, 887–892. doi: 10.1073/pnas.1418031112

Bergson, H. (1911). *Matter and Memory*. London: Allen and Unwin.

Bhattacharya, A., Aghayeva, U., Berghoff, E. G., and Hobert, O. (2019). Plasticity of the electrical connectome of C. elegans. *Cell* 176, 1174–1189. doi: 10.1016/j.cell.2018.12.024

Brown, T. (2003). *Making Truth: Metaphor in Science*. Chicago: University of Illinois Press.

Bullock, T. H., and Horridge, A. (1965). *Structure and Function in the Nervous Systems of Invertebrates*. San Francisco: W. H. Freeman.

Butcher, W. D. (1912). The education of the brain as an electric organ. *J. Rönt. Soc.* 8, 76–90. doi: 10.1259/jrs.1912.0046

Cajal, S. (1894). The Croonian lecture – la fine structure des centres nerveux. *Proc. Roy. Soc. Lond.* 55, 444–468. doi: 10.1098/rspl.1894.0063

Cazé, R., Humphries, M., and Gutkin, B. (2013). Passive dendrites enable single neurons to compute linearly non-separable functions. *PLoS Comp. Biol.* 9:e1002867. doi: 10.1371/journal.pcbi.1002867

Churchland, A., and Abbott, L. (2016). Conceptual and technical advances define a key moment for theoretical neuroscience. *Nat. Neurosci.* 19, 348–349. doi: 10.1038/nn.4255

Cobb, M. (2020). *The Idea of the Brain: a History*. London: Profile.

Cole, K. S., and Curtis, H. J. (1939). Electric impedance of the squid giant axon during activity. *J. Gen. Physiol.* 22, 649–670. doi: 10.1085/jgp.22.5.649

Crick, F. H. C., and Jones, E. (1993). Backwardness of human neuroanatomy. *Nature* 361, 109–110. doi: 10.1038/361109a0

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering how the Brain Codes Our Thoughts*. New York: Penguin.

Frégnac, Y. (2017). Big data and the industrialization of neuroscience: a safe roadmap for understanding the brain? *Science* 358, 470–477. doi: 10.1126/science.aan8866

Friston, K., Moran, R., and Seth, A. K. (2013). Analysing connectivity with granger causality and dynamic causal modelling. *Curr. Opin. Neurobiol.* 23, 172–178. doi: 10.1016/j.conb.2012.11.010

Gomez-Marin, A. (2021). Promisomics and the short-circuiting of mind. *eNeuro* 8:ENEURO.0521-20.2021. doi: 10.1523/ENEURO.0521-20.2021

Hagmann, P. (2005). *From Diffusion MRI to Brain Connectomics*. PhD Thesis. Lausanne: ePFl.

Heimer, L. (1971). Pathways in the brain. *Sci. Am.* 225, 48–64. doi: 10.1038/scientificamerican0771-48

Helmholtz, H. (1875). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. London: Longmans, Green. doi: 10.1037/10838-000

Kandel, E. R. (1970). Nerve cells and behaviour. *Sci. Am.* 223, 57–71. doi: 10.1038/scientificamerican0770-57

Lakoff, G., and Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

McCulloch, W. S., and Pfeiffer, J. (1949). Of digital computers called brains. *Sci. Monthly* 69, 368–376.

McCulloch, W., and Pitts, W. (1943). The immanent logic of the nervous system. *Bull. Math. Biophys.* 5, 115–133. doi: 10.1007/BF02478259

Meyerott, R. E. (1946). Mathematical biophysics of the central nervous system. *Am. J. Sci.* 244, 865–866.

Pitts, W. (1955). "Comments on session on learning machines," in *Proceedings of the March 1-3, 1955, Western Joint Computer Conference, 108-111*. New York, NY: Association for Computing Machinery doi: 10.1145/1455292.1455313

Pribram, K. (1969). The neurophysiology of remembering. *Sci. Am.* 220, 73–86. doi: 10.1038/scientificamerican0169-73

Robinson, J. (2001). *Mechanisms of Synaptic Transmission: Bridging the Gaps (1890–1990)*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195137613.001.0001

Rosenblatt, F. (1958). The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386–408. doi: 10.1037/h0042519

Smee, A. (1850). *Instinct and Reason Deduced from Electro-Biology*. London: Reeve, Benham and Reeve. doi: 10.1037/12035-000

Sporns, O. (2015). "Network neuroscience," in *The Future of the Brain: Essays by the World's Leading Neuroscientists*, eds G. Marcus and J. Freeman (Oxford: Princeton University Press), 90–99. doi: 10.1515/9781400851935-012

Sporns, O., Tononi, G., and Kötter, R. (2005). The human connectome: a structural description of the human brain. *PLoS Comp. Biol.* 1:e42. doi: 10.1371/journal.pcbi.0010042

Stadler, M. (2017). "Circuits, algae, and whipped cream: the biophysics of nerve, ca. 1930," in *The History of the Brain and Mind Sciences: Technique, Technology, Therapy*, eds S. T. Casper and D. Gavrus (Rochester NY: Boydell & Brewer). doi: 10.1017/9781787440159.005

Stern, S., Kirst, C., and Bargmann, C. I. (2017). Neuromodulatory control of long-term behavioral pattenrs and individuality across development. *Cell* 171, 1649–1662. doi: 10.1016/j.cell.2017.10.041

Sutherland, N. S. (1974). "Computer simulation of brain function," in *Philosophy of Psychology*, ed. S. C. Brown (London: Macmillan Education). doi: 10.1007/978-1-349-02110-9_13

Swammerdam, J. (1758). *The Book of Nature*. London: Seyffert.

Thomson, S., and Smith, H. (1853). *A Dictionary of Domestic Medicine and Household Surgery*. Philadelphia: Lippincott.

Troland, L. T. (1922). The present status of visual science. *Bull. Nat. Res. Coun.* 5, 1–120.

White, J. G. (2013). "Getting into the mind of a worm—a personal view (June 25, 2013)," in *WormBook*, ed. The *C. elegans* Research Community. doi: 10.1895/wormbook.1.158.1

White, J. G., Southgate, E., Thomson, J. N., and Brenner, S. (1986). The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philos. Trans. R. Soc. Lond. B* 314, 1–340. doi: 10.1523/jneurosci.11-01-00001.1991

Witvliet, D., Mulcachy, B., Mitchell, J. K., Meirovitch, Y., Merger, D. R., Wu, Y., et al. (2021). Connectomes across development reveal principles of brain maturation. *Nature* 596, 257–261. doi: 10.1038/s41586-021-03778-8

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

GENERAL COMMENTARY
published: 29 June 2022
doi: 10.3389/fcomp.2022.890531

Commentary: Metaphors We Live By

*Alex Gomez-Marin**

Instituto de Neurociencias (CSIC-UMH), Alicante, Spain

Keywords: neuroscience, computer, metaphor, thought, productive vs. permissive brain function

A Commentary on

Metaphors We Live By
Lakoff, G. and Johnson, G. (2003). Chicago, IL: The University of Chicago Press, p. 276.

Four decades ago, linguists and philosophers George Lakoff and Mark Johnson published an influential book on the nature of metaphors. In *Metaphors We Live By* they argued that abstract thought is mostly metaphorical (having a literal core extended by mutually inconsistent metaphors and therefore incomplete without them), that metaphors are fundamentally conceptual (while metaphorical language is secondary), and that metaphorical thought is ubiquitous, unavoidable, largely unconscious, and grounded in everyday life. Despite the popular acclaim of the book and its impact across academic disciplines, their claims met resistance as they challenged objectivist views of meaning and language.

Two decades ago, in the afterword of the updated edition of their classic book, the authors insisted on some persistent fallacies that contribute to a false view of what metaphors are and do. Especially relevant to scientists is the mistaken idea that metaphors are just "a matter of words," namely, a way of speaking that can be shielded from rational thinking, and ultimately innocuous to it. Such a fallacy is related to the belief that metaphors are mainly "a matter of definition;" "it is just semantics" is stupid as a conversation stopper. After all, meaning is all that matters. Furthermore, Lakoff and Johnson remarked that metaphors are natural phenomena, not mere arbitrary historical contingencies or cultural constructions; they are grounded in the very bodily nature of our daily cognitive pursuits.

Today, neuroscientists remain ensnared in disputes about whether brains are computers or not. The debate about the appropriateness and obsolescence of the brain-computer analogy and its pervading metaphorical use is alive and ticking. Of course, some will deem the exchange as useless, while others will insist that, until we figure out what we mean by what we say, a proper scientific discussion is defective or cannot even start.

Here, rather than poking the brain-as-computer blister again, I would like to walk through the very brief history of the metaphorical brain, and then mention an alternative class of metaphors that could offer fresh, offbeat, and even rebellious perspectives on our conception of that mushy little thing called "the brain."

Brains have been many things indeed. Four centuries ago, Descartes suggested that cerebral hydraulic automata produced behavior by powering "animal spirits" thought the nerves. Nicolas Steno cast the brain as a machine. Fountain metaphors gave way to clocks. Leibniz objected: entering a brain as one enters a mill would reveal mechanical parts but nothing mind-like whatsoever. In the time of Shelley's *Frankenstein*, Galvani, and Volta explored the role of electricity in animal bodies. Nerves turned into wires and brains into telegraphs. Neural plasticity soon hindered the analogy. Cajal preferred natural images: trees, gardens, forests. For Darwin, thought was a "secretion" of the brain. It was soon postulated that all animals, including humans, were "conscious machines." But machines would become animal-like too: electric dogs, clockwork

Edited by:
Pedro Martinez,
University of Barcelona, Spain

Reviewed by:
Daniel J. Graham,
Hobart and William Smith Colleges,
United States

****Correspondence:***
Alex Gomez-Marin
agomezmarin@gmail.com

Specialty section:
This article was submitted to
Theoretical Computer Science,
a section of the journal
Frontiers in Computer Science

Received: *06 March 2022*
Accepted: *22 April 2022*
Published: *29 June 2022*

Citation:
Gomez-Marin A (2022) Commentary:
Metaphors We Live By.
Front. Comput. Sci. 4:890531.
doi: 10.3389/fcomp.2022.890531

beetles, three-wheeled moths. With the theoretical articulation and empirical implementation of feedback loops, the line between biology and technology thinned. Pitts and McCulloch's "neural network" further blurred the distinction between the natural and the artificial. The metaphor was turned inside-out: computers were like brains, then brains became computers. With information theory in place, code-breaking percolated from the war to the lab. Physicalists stopped worshiping matter and revered information instead, a notion as pervasive as confounded. The brain has, thenceforth, been deemed as the preeminent computational organ. Despite considerable progress and obvious updates of the metaphor *du jour* (networks, the internet, and so on), we have hardly gone much further.

The less told story of the idea of the brain involves other images. A remarkably forgotten one, inspired by a fascinating physical technique, is the hologram. The brain would be a holographic information device. Modulating waves within a field, a part could contain the whole, allowing for non-local memory storage (the still elusive engram would be lost and found). Simple metaphors can also be conceptually juicy. Consider a prism, whereby light is reflected (perception as virtual action) and refracted (embodiment and affect). Blue would not be created in the prism but selected from the incoming beam. As Henri Bergson put it, the brain would be the organ of attention to life, whose main role is to receive, delay, and conduit movement, carving out external images rather than producing them. From this perspective, brains are more like radios than VR headsets. This whole class of metaphors leads us back to William James' foundational (and forgotten) distinction between brain function as "productive" vs. "permissive." The brain as a "reducing valve," in Aldous Huxley's words, is an intriguing hypothesis that could be investigated within the current renaissance of psychedelic research. Brains would not create thoughts but receive and filter them. These are other metaphors neuroscientists could live by.

It is also helpful to note that English is not the only language spoken by human beings. The word *computer* is not universally translated as such. A computer is still a "computer" in Italian, Portuguese, or German. But in Catalan we say *ordinador*, which we took from the French *ordinateur*. The same happens in Spanish with *ordenador* (except outside of Spain, where people say *computadora*). The story behind the choice of words is peculiar. In 1955, the IBM marketing team in France decided against branding their new product too similarly to existing "calculators." As a better (and shorter) name for "the new electronic (programable) machine intended for information processing," they decided on *ordinatrice électronique*. Finally, *ordinateur* settled as a trademark, percolating to current language. Depending on the country, brains are pictured in the image of a person who organizes or who computes. We can remain lost in translation.

Like fish in the sea, we often fail to notice the entrenched metaphors we swim in. Brains are not really *any* of those, and yet treating brains *as such* can provide valuable insights *unless* one does not erect one's favorite image into an idol. As Lewontin's quote of Wiener and Rosenblueth puts it, "the price of metaphor is eternal vigilance." A metaphorical monoculture is a burden rather than a blessing. Glossing George Box's aphorism, all metaphors are wrong (when literally taken), but all are useful (when kept in their local domain of application). Screwdrivers are handy, but not to eat soup. Entertaining other ways to conceive what brains are, and what they do, is not only valuable but necessary.

Moreover, in the light of Lakoff and Johnson's thesis, honoring the metaphorical nature of much of our scientific thinking frees our imagination and allows us to deliberately explore the many-sided nature of the brain. In the bigger picture, *Metaphors We Live By* was a reaction to the tendency within the analytical tradition to demote metaphors as either meaningless or simply pragmatic vectors to literal meanings. The thesis of the book was discordant with those provinces of the Western canon that ascribe an immaculate purity to concepts, but also with the skeptic relativism of postmodernist doctrines. The metaphors we use shape what we can and cannot *see*, both under our microscopes but also in the real world. When it comes to understanding human mental life, the study of metaphors is complementary to the study of brains themselves. As we think, we live.

AUTHOR CONTRIBUTIONS

AG-M wrote the manuscript.

FUNDING

AG-M was supported by the Spanish Ministry of Science.

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Ecology and Evolution

TYPE Review
PUBLISHED 09 August 2022
DOI 10.3389/fevo.2022.952992

Artifacts and levels of abstraction

M. Chirimuuta*

Department of Philosophy, University of Edinburgh, Edinburgh, United Kingdom

OPEN ACCESS

EDITED BY
Pedro Martinez,
University of Barcelona, Spain

REVIEWED BY
Javier Suárez,
Jagiellonian University, Poland
Denis Walsh,
University of Toronto, Canada

*CORRESPONDENCE
M. Chirimuuta
m.chirimuuta@ed.ac.uk

SPECIALTY SECTION
This article was submitted to Models in Ecology and Evolution, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 25 May 2022
ACCEPTED 08 July 2022
PUBLISHED 09 August 2022

CITATION
Chirimuuta M (2022) Artifacts and levels of abstraction. *Front. Ecol. Evol.* 10:952992. doi: 10.3389/fevo.2022.952992

The purpose of this article is to show how the comparison or analogy with artifacts (i.e., systems engineered by humans) is foundational for the idea that complex neuro-cognitive systems are amenable to explanation at distinct levels, which is a central simplifying strategy for modeling the brain. The most salient source of analogy is of course the digital computer, but I will discuss how some more general comparisons with the processes of design and engineering also play a significant role. I will show how the analogies, and the subsequent notion of a distinct computational level, have engendered common ideas about how safely to abstract away from the complexity of concrete neural systems, yielding explanations of how neural processes give rise to cognitive functions. I also raise worries about the limitations of these explanations, due to neglected differences between the human-made devices and biological organs.

KEYWORDS

philosophy of neuroscience, levels of abstraction, levels of explanation, analogy, philosophy of cognitive science

Introduction

It is worth remembering that the very word *organism* comes to us via an analogical transfer from the Greek word for tool (*organon*), and originally meant the property of things comprising heterogenous parts that work together in a coordinated way—a property pretty much captured by the word *mechanism* today (Cheung, 2006; Illetterati, 2014, p. 89). What this indicates is that even when drawing contrasts between organisms and machines, organs and artifacts, the concepts we are using to theorize living beings have originated through a process of comparison with objects that people have made. As philosopher Martin Heidegger observed, "[p]erhaps it will take a long time to realize that the idea of organism and of organic is a purely modern, mechanical-technical concept, so that what grows naturally by itself is interpreted as an artifact that produces itself" (quoted in Nunziante, 2020, p. 12).

This special issue invites us to weigh up the claim that *all metaphors are false but some are useful*. Incidentally, our notion of the useful, utility, is shaped by concept of the *tool*—the tool is the paradigmatic useful thing. This connection is made obvious in the French language, where the words for tool (*outil*) and useful (*utile*) are so similar.

Thus, we have on the one hand, the root metaphor of the living being (*organism*) as a system of tool-like components (*organs*), and on the other, the question of whether metaphors gathered from the making and employment of tools and machines is itself a useful *conceptual tool*.

This essay will zoom in on one aspect of the comparison between immensely complex nervous systems, and relatively simple information processing machines: the idea that the brain, like the computer, can be explained at distinct and somewhat autonomous levels of analysis. I will account for the utility of this analogy as due to its providing a simplifying strategy for neuroscientists. The assumption that there is a "high level" description of the brain which can be modeled and comprehended in the absence of detailed knowledge of the "low level" components is motivated by consideration of the hardware/software distinction in computing. I will illustrate this strategy via an exposition of David Marr's well known system of levels of explanation (section "Marr's levels of explanation"). We will then see how the levels framework is motivated by analogies with machines, primarily computers, but also with the procedures that people undertake when making devices, here analyzing the influential ideas of Herbert Simon on hierarchical complex systems (section "The artifact analogies"). A risk of reliance on such analogies is that it leads to neglect of differences. All analogies are imperfect, but sometimes researchers forget this. The section "Limitations of the analogies" considers the limitations of the analogy between messy "heterarchical" biological systems and man-made designs that have a clearly delineated modularized and leveled structure. To conclude, I ask whether these limitations can be addressed through comparison with more life-like machines—as suggested in this special issue by Bongard and Levin (2021). I argue that their proposal neglects the problem of opacity that comes with the introduction of more complex machine models.

Immauel Kant is one philosopher whose account of biological knowledge recognized that the comparison between the workings of nature, and processes of engineering was indispensable to the conceptualization of living beings: both biology and engineering rely on *functional* notions, the understanding of certain processes happening for the sake of wider system-level goals. At the same time, he warned against an anthropomorphism that comes with taking this as the literal, ultimate truth about the natural world. He wrote in the *Critique of the Power of Judgment* that, "we picture to ourselves the possibility of the [biological] object on the analogy of a causality of this kind—a causality such as we experience in ourselves—and so regard nature as possessed of a capacity of its own for acting *technically*" (Kant, 1790/1952, Part II p.5/§361; Breitenbach, 2014). But as Illetterati (2014, p. 91) explains, "these kinds of notions, even if necessary, seem to maintain a sort of fictional character too: indeed, they have no justification in things themselves, but neither do they have their origin in mere human invention. They rather have their justification in the way subjects necessarily understand living beings." I think that this is the right way to interpret machine analogies in biology, and engineering metaphors more generally: they are useful precisely because they allow scientists to figure nature in human terms, which is why they are—strictly speaking—false.

Marr's levels of explanation

As is well known, Marr's framework is introduced in the first chapter of his book, *Vision* (Marr, 1982, p. 25). The three levels are:

1. Computational theory
2. Representation and algorithm
3. Hardware implementation

The "top level" computational theory gives an abstract characterization of the performance of a system in terms of its generating a mapping of an input to an output. In addition, characterization at this level shows how that performance is related to environmental constraints and behavioral goals. Thus, the first level is to provide a functional characterization in both senses of the word: explicating a mathematical input-output mapping, and also illuminating the utility of the performance. The middle level involves specification of the format for representation of the inputs and outputs, and of the algorithm that transforms one into the other. The bottom level describes how the representations and algorithm are physically realized, for example in the electronic components of a computer vision system, or in the neurons of an animal's retina.

In the next section I will say more about how analogies with machines motivate this three-level system, and why they are essential in the interpretation of it. Here we should note that Marr's proposal carries on from a discussion of the limitations of reductionist approaches to explaining the visual system—attempts to understand how neural activity gives rise to useful perceptions of the environment by way of careful study of the anatomy and physiology of neurons. In effect, the reductionist is restricted to the bottom level of explanation. Marr (1982, p. 27) describes this approach as equivalent in futility with the attempt to understand bird flight just through the examination of feathers. As he asserts in the preamble to the three levels, "[a]lmost never can a complex system of any kind be understood as a simple extrapolation from the properties of its elementary components" (Marr, 1982, p. 19). The basic complaint against reductionism is that this is a strategy that quickly gets the investigator overwhelmed with details whose significance cannot be assessed because she lacks knowledge of the overall functionality of the system, and therefore has no working hypothesis about how the elementary components contribute to global properties and behavior. The shape of the forest is invisible because there are so very many leaves. The introduction of the two additional levels of explanation allows for lines of investigation that prioritize general questions about the system's functionality and operations independently of investigation into implementational details. The upper two levels are *levels of abstraction* away from the concrete, complicated material system. Ideally, the results of these upper level investigations provide a map of what to look for in the concrete system, and a guide to interpreting the material details, even though the levels are only "loosely related" (Marr, 1982, p. 25).

One of the virtues of Marr's framework, highlighted by later researchers, is that it offers this strategy for simplification.[1] For example, Ballard (2015, p. 13) writes that it, "opened up thinking about the brain's computation in abstract algorithmic terms while postponing the reconciliation with biological structures." Speaking of level schemas more generally, Ballard emphasizes that, "[b]y telescoping through different levels, we can parcellate the brain's enormous complexity into manageable levels" (2015, p. 18).

The artifact analogies

The general impression given by Marr's presentation is that he does not care to set a division between engineered and living systems, between those that have (computational) functions, properly speaking, and those for which it is only a heuristic posit. A striking feature of Marr's presentation is that in the first instance it relies exclusively on examples of information processing machines. Cases from within neuroscience are mentioned only after a complete account of the three levels has been given, without there being any comment on this transition. The primary illustration of the levels comes by way of a cash register, an adding machine. At the computational level, the task is to find out "*what* the device does and *why*." (Marr, 1982, p. 22).[2] This means specification of the arithmetical theory of addition, as well as an account of the functional role of the machine for adding up charges in a shop. We learn that the second level characterization involves showing how numbers are represented in the device (e.g., Arabic or Roman notation), and specifying the algorithm used to work out the total bill. The implementation level involves characterization of the "physical substrate" which runs the algorithm. A point Marr (1982, p. 24) emphasizes is that the same algorithm can be realized in very different materials. This also goes for the relationship between the top two levels: one and the same computational task can be achieved by a range of different algorithms. This is why the levels are only "loosely related" (p. 25)—a discovery at one level cannot reliably pre-specify what will be found at the level below.

We might speculate that Marr leans on artifacts for purposes of exposition just because the core concept of each of these levels comes out especially clearly in cases like the cash register. But then we ought to wonder why it is that it is harder to get a grip on how to define these levels in neuroscience, even though the framework is intended for use there. We can discern a deeper reason for the primacy of machines in Marr's exposition if we consider Dennett's observation that the three levels actually schematize the stages taken in the engineering of a complex information processing system. Dennett (1995, p. 682) writes,

> Marr's obiter dicta [passing words] on methodology gave compact and influential expression to what were already reigning assumptions in Artificial Intelligence. If AI is considered as primarily an engineering discipline, whose goal is to create intelligent robots or thinking machines, then it is quite obvious that standard engineering principles should guide the research activity:

> first you try to describe, as generally as possible, the capacities or competences you want to design,

> and then you try to specify, at an abstract level, how you would implement these capacities,

> and then, with these design parameters tentatively or defeasibly fixed, you proceed to the nitty-gritty of physical realization.

The point here is that the three levels of explanation are an expression of three broad steps in the *forward engineering* of a machine with some functionality equivalent to a cognitive capacity in an animal. It is then not surprising that the different

[1] Of course, the details of Marr's framework have been criticized by later researchers, such as Love (2021), who argue for a greater number of levels. Gurney (2009) proposes a four-level framework which is incidentally more similar to one proposed by Marr in a 1976 technical report.

[2] To reinforce this point about the primacy of artifacts, note that Marr does not use the neutral language of "things" or "systems" but refers specifically to a "device" here. We find this also in the legend for the summary table: "The three levels at which any *machine* carrying out an information-processing task must be understood" (p. 25 emphasis added). Cf. "the different levels at which an information processing *device* must be understood before one can be said to have understood it completely" (p. 24 emphasis added).
Later in the book, when again summarizing the three levels as applied to the visual system, it is interesting that the terms "machine" and "machinery" are still used:
"The human system is a working example of a machine that can make such descriptions, and as we have seen, one of our aims is to understand it thoroughly, at all levels: What kind of information does the human visual system represent, what kind of computations does it perform to obtain this information, and why? How does it represent this information, and how are the computations performed and with what algorithms? Once these questions have been answered, we can finally ask, How are these specific representations and algorithms implemented in neural machinery?" (Marr, 1982, p. 99).

levels are more easy to illustrate with an example of *reverse engineering* some such device.

The issue I am highlighting here is that artifacts are the foundational cases for Marr's framework, and the application to neuroscience occurs via an analogical transfer to brains, systems which are arguably similar to computing ones. Researchers habitually think of brains, just like the artifacts, as taking in inputs (e.g., from sensory organs), implementing some algorithms, and sending an output (e.g., a motor command).[3] The importance of this analogy comes out in Dennett's characterization of what his own approach has in common with that of Marr and cognitive scientist Allen Newell, namely:

> stress on being able (in principle) to specify the function computed . . .. independently of the other levels.
>
> an optimistic assumption of a specific sort of functionalism: one that presupposes that the concept of the function of a particular cognitive system or subsystem can be specified (It is the function which is to be optimally implemented.)
>
> A willingness to view psychology or cognitive science as reverse engineering in a rather straightforward way. Reverse engineering is just what the term implies: the interpretation of an already existing artifact by an analysis of the design considerations that must have governed its creation (Dennett, 1995, p. 683).

Dennett's articulation of the reverse engineering methodology, his *design stance,* comes with strict assumptions of optimality and adaptationism in evolved systems that we need not attribute to the scientific practice. In my view, the essential point about the reverse engineering methodology is that it treats the biological object by analogy with a man-made thing, and in this way attempts to make it intelligible by showing how it operates according to principles that make sense from the perspective of a person designing things; in other words, by treating it as if it were an artifact, the scientist can explain it in terms of the practical rationality of causal means being used to produce useful effects.

We should appreciate that there are two levels of analogy, so to speak. Superficially, the analogy just holds between certain organs of living bodies and man-made devices that have a rough functional equivalence with them—the brain and a computer, the heart and a pump. But the deeper and more general point—the one spelled out by Kant—is that there is an analogy being invoked between the systematic organization of parts and processes through which organs generate their functional effects, and the parts and processes set in place by a human engineer in order for a device to achieve the desired effect. An artifact is intelligible to the extent that its operations are the manifestations of the instrumental rationality through which its human makers have put components together in order to achieve their goals. A similar kind of intelligibility is tacitly assumed for the biological object. This becomes clearer when we consider *functional analysis*, which is a general schema for reverse engineering (see **Figure 1**).

The link between this reverse engineering methodology in cognitive and neuroscience, and simplification of the brain becomes apparent if we focus on the importance of encapsulation in functional analysis. When a system is described in this way, the payoff is that at any given level of analysis the component modules can be treated as black boxes whose inner workings are either unknown or ignored, since the only information relevant to the current level of analysis is the input-output profiles of the modules. Descent to a lower level of

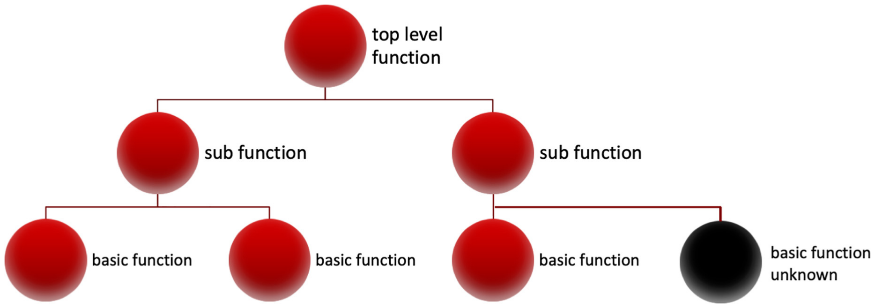


FIGURE 1
Reverse engineering is expressed schematically as performance of a functional analysis (Cummins, 1975, Cummins, 2000). The top level function of the whole system is decomposed into sub-functions, which can themselves be explained in terms of the interaction of basic functions. See also Bechtel and Richardson (2010) on the research strategy of functional decomposition, employed for investigating modular, hierarchical systems.

[3] E.g., Marcus and Freeman (2015, p. xiii):
"The brain is not a laptop, but presumably it is an information processor of some kind, taking in inputs from the world and transforming them into models of the world and instructions to the motor systems that control our bodies and our voices."
See Chirimuuta (2021, under contract) on why this practice should be interpreted as resting on a loose analogy rather than strict functional similarity between computer and brain.

analysis involves opening the black boxes and seeing how their inner workings can be accounted for in terms of the functional capacities of their components. But for many explanatory purposes, lower level details can safely be kept out of view, which is why this methodology offers a handy simplification.

To illustrate this point, I will make use of an example from computing given by Ballard (2015, p. 14ff.). Most people who program computers only ever use a high level programming language such as Python. But the terms of this high level language are actually black boxes which unpack into more complicated expressions in a lower level assembly language. These lower level terms themselves unpack into instructions in machine code. For a program to be carried out, it needs to be translated down into lower level languages, "closer to machine's architecture." But this is all done behind the scenes and the ordinary coder can comfortably stick with description of the computation in the compact, highest level language. The point of Ballard's example here is to argue that there is a tight analogy between the computer and the brain, which he thinks can be described similarly in terms of "levels of computational abstraction."

Crucially, the abstraction hierarchy is posited to be there in the brain's own representations of the extra-cranial world, not just in those imposed upon it by a scientist. The proposal is that the brain is a system that, at the top level of control, ignores its own complexity, like a digital computer where the execution of a piece of code is indifferent to micro-physical fluctuations in the electronic hardware. Just as the programmer, the controller of a computer, can govern the performances of the machine while ignoring and remaining ignorant of its low-level languages and physical workings, it is supposed that the brain systems ultimately responsible for behavior employ an abstract, high level system of representation that is invariant to changes in the complex, low-level workings of the brain and rest of the body. If this assumption holds, there are good prospects for a relatively simple computational theory that explains how the brain governs behavior, by way of these high-level representations.

But why would neuroscientists think that this assumption does hold, that the analogy between computer and brain is tight enough? The intelligible organization of systems as hierarchically arranged, encapsulated modules, or levels of more or less abstract representations, can be found in artifacts designed by humans, but its existence in the natural world should not be taken for granted. As far as I can determine, the foundational argument in support of this assumption comes from another analogy put forward by Herbert Simon in the "Architecture of Complexity" (Simon, 1962, 1969).[4] In a tale of two watchmakers, Simon describes how the production of a complex system (a watch) is much more likely to be successful if the production process occurs in stages, where sub-processes in the production result in stable sub-components of the system that are assembled together at a later stage. Simon then draws an analogy between human manufacture and the evolution of complex life forms. His point is that the likelihood of evolution producing organisms of any complexity is vanishingly small unless it is the case that it comes about via the evolution of intermediate, self-standing forms that become the components of more complex organisms. Hence, he argues, it must be the case that evolved, as well as manufactured complex systems are composed, hierarchically, of relatively independent sub-systems. In these *near decomposable* complex systems, there is only a weak frequency and strength of interaction horizontally between the subsystems at any one level, and vertically across the levels of organization. This means that the subsystems—the modular components—can usefully be studied in isolation from the rest of the system, and that the system can be studied at higher levels of organization (which we can here equate to larger scales) without attention to most of the lower level (i.e., small scale) details. The optimistic upshot is that evolved complex systems are scientifically intelligible through decomposition into levels and components, and that this is an alternative to intractable reductionist methodologies.[5]

Reductionist methodologies can be successful for relatively simple systems. The task of the research is to acquire sufficient information about the elementary components, and their interaction, to yield an explanation of the behavior of the whole system. This is a "flat," as opposed to multi-level, approach. Once there is enough complexity that the amount of information about elementary components and the interactions that can feasibly be dealt with (in models or theory) is much less than what is required for explanation of the system's behavior, then a multi-level approach is needed. The common virtue of all of the multi-level approaches discussed above—from Marr, Ballard, and Simon—is that they offer a guide for how to abstract away from low-level details and how to set about work on top-down explanations when

[4] It is interesting that Marr (1982, p. 102) also makes the connection between evolvability, intelligibility, and modular organization:
"This observation [of isolated visual processing] .... is fundamental to our approach, for it enables us to begin separating the visual process into pieces that can be understood individually. Computer scientists call the separate pieces of a process its *modules*, and the idea that a large computation can be split up and implemented as a collection of parts that are as nearly independent of one another as the overall task allows, is so important that I was moved to elevate it to a principle, the *principle of modular design*. This principle is important because if a process is not designed in this way, a small change in one place has consequences in many other places. As a result, the process as a whole is extremely difficult to debug or to improve, whether by a human designer or in the course of natural evolution, because a small change to improve one part has to be accompanied by many simultaneous, compensatory changes elsewhere. The principle of modular design does not forbid weak interactions between different modules in a task, but it does insist that the overall organization must, to a first approximation, be modular."

[5] See Bechtel and Richardson (2010) for further discussion of methods for investigating near decomposable systems.

bottom-up, reductionist approaches are intractable, even if possible in principle. These three scientists are all advocates of computational explanations of how the brain gives rise to cognition, and this kind of explanation is favored because, they argue, it does not require that much attention be paid to the details of neurophysiology which would otherwise threaten an overwhelming complexity.

An additional feature of computational explanations is that they assert an equivalence between organic and artificial systems, so long as they are computing the same functions. This is known in philosophy as *multiple-realization*. A mechanical cash register, an electronic calculator, and a human brain region, can all be said to be doing the same computation when adding up a particular sum, even though the physical substrates are so different. The benefit of this for neuroscientific research is that it justifies the substitution of actual neural tissue with *relatively* simple computational models, such as artificial neural networks (ANNs), as objects of investigation. A goal of various neuro-computational research projects has been to create models of brain areas *in silico* that will yield confirmatory or disconfirmatory evidence for theories of cognition and pathology, where traditional experimental approaches are untenable because it is not possible to make the required interventions on actual neurons. Even though large ANNs are themselves rather complicated and hard to interpret, they are at least more accessible to (simulated) experimental interventions, such as lessoning of individual nodes.

Aside from the specifics of computational explanation (explanation via analogy between brains and computers), one of the general implications of the artifact analogy is that the nervous system is composed of relatively encapsulated working parts (modules) or functional components. This also supports the "black-boxing" of neural details. As Haugeland (1978, p. 221) relates,

> if neurons are to be functional components in a physiological system, then some specific few of their countless physical, chemical, and biological interactions must encapsulate all that is relevant to understanding whatever ability of that system is being explained.

One way to think about the importance of *neuron doctrines* in the history of the discipline—theories that posit individual neurons as the basic anatomical and functional units of the nervous system—is that they facilitate this simplifying strategy, even while departing from many of the observable results on the significance of sub-neuronal and non-neuronal structures and interactions.[6] Moreover, we should note also that this black-boxing can be employed to achieve abstract representations of functional components other than individual neurons (e.g., Hawkins et al., 2017 model of cortical columns).

Limitations of the analogies

I have argued that the dominant multi-level approaches in neuroscience rest on the assertion of there being a close similarity between the multi-level organization of artifacts such as computers, and the brain, an evolved organ whose organizational "plan" is far less well characterized than that of the machine, and remains a matter of controversy. This prompts consideration of the difficulties that the multi-level approach faces, to the extent that the claim for similarity can be challenged. If the comparison between brain and computer is at best a loose analogy, in which the dissimilarities between the two are of equal importance or even outnumber the similarities, then the leveled approach might sometimes be a hindrance in the project of explaining how brain activity gives rise to cognition.

The first concern to bring up here is that the case for encapsulation in the nervous system is fairly weak. This was pointed out decades ago by Haugeland, in the passage following on from the one quoted above:

> [encapsulation] is not at all guaranteed by the fact that cell membranes provide an anatomically conspicuous gerrymandering of the brain. More important, however, even if neurons were components in some system, that still would not guarantee the possibility of "building back up." Not every contiguous collection of components constitutes a single component in a higher level system; consolidation into a single higher component requires a further encapsulation of what's relevant into a few specific abilities and interactions—usually different in kind from those of any of the smaller components. Thus the tuner, pre-amp and power amp of a radio have very narrowly specified abilities and interactions, compared to those of some arbitrary connected collection of resistors, capacitors, and transistors. The bare existence of functionally organized neurons would not guarantee that such higher level consolidations were possible. Moreover, this failure of a guarantee would occur again and again at every level on every dimension. There is no way to know whether these explanatory consolidations from below are possible, without already knowing whether the corresponding systematic explanations and reductions from above are possible—which is the original circularity (Haugeland, 1978, p. 221).

It is interesting that Haugeland focuses on the possibility of a strong disanalogy between the organization of the nervous

[6] See Bullock et al. (2005) and Cao (2014) on the empirical inadequacy of the neuron doctrine. Barlow (1972) is a great example of its role in explanatory simplification.

system, and that of a human-designed artifact, a radio. Whereas it is a feature of the design of a radio that higher level sub-components (the tuner, pre-amp and power amp) are made up of careful arrangements of lower level sub-components (resistors, capacitors, and transistors), and themselves have narrowly specified capacities and input-output profiles, it should not be assumed that collections of neurons consolidate into higher level sub-components in this way, and that explanations of the neural basis of cognition can safely be restricted to the higher levels. I will now discuss two reasons to be skeptical that the analogy holds. The first relates to the potential importance of low-level activity, the second brings up the difference between hierarchical, designed systems and evolved ones.

It is an open possibility that cognition is the product of dense interactions across a number of levels or scales, and is not restricted to a high level of computational abstraction, as hypothesized by Ballard. The cognitive properties of the brain may be enmeshed in its material details, in a way not congenial to Marr's vision of a there being computational and algorithmic/representation levels that are only loosely related to the implementational one. A reason to give credence to these possibilities comes from consideration of the fact that biological signaling, a general feature of living cells, is the omnipresent background to neuronal functionality. The low level details of neuronal activity can themselves be characterized as doing information processing, and are not merely the hardware implementors of the system's global computations, or bits of infrastructure keeping the system running. This is an argument put forward by Godfrey-Smith (2016, p. 503):

> This coarse-grained cognitive profile is part of what a living system has, but it also has fine-grained functional properties—a host of micro-computational activities in each cell, signal-like interactions between cells, self-maintenance and control of boundaries, and so on. Those finer-grained features are not merely ways of realizing the cognitive profile of the system. They matter in ways that can independently be identified as cognitively important.

The point is that in an electronic computer there is a clean separation of the properties of the physical components that are there holding the device together, and the ones involved specifically in information processing. This is how the machine has been designed. Whereas in the brain this is not the case—it is not clear cut which entities within the brain, and which of their properties, are responsible for information processing, and which are the infrastructural background.[7] In addition to the "coarse-grained" computations that might be attributed on the basis of the whole animal's psychology and behavior, Godfrey-Smith argues that there are a countless number of "micro-computational activities" in cells, which are not unrelated to global cognition. If in the brain metabolism, cell-maintenance, and global (i.e., person-level) cognitive functions are enmeshed together, then low level material details about neural tissue, such as the specific chemical structures of the many kinds of neurotransmitter, and the thousands of proteins expressed at synapses (Grant, 2018), probably do matter to the explanation of cognition. They cannot be safely discounted with the same confidence as merited in aeronautics, when air is treated as a continuous fluid and molecular details are left unrepresented.[8]

We saw that Herbert Simon gives an in principle argument for the existence of hierarchical organization in complex living systems which would, if accepted, justify the exclusion of low level details for the purposes of most explanations of whole system behavior. However, the strict analogy this argument supposes, between human manufacture and the processes of evolution, calls for scrutiny. Bechtel and Bich (2021) argue that hierarchical control structures, with their neat pyramidal arrangement of superordinate and subordinate levels, are less likely to evolve than *heterarchical systems*, which have a more haphazard arrangement of horizontal and vertical interconnections, meaning that one component of the system is open to significant influence from components at other levels (they are not just "loosely related"), and there is no top-level locus of control, as posited by Ballard (2015, p. 242) in his comparison between control in robots and humans. The reason for the hypothesized predominance of heterarchical systems is that evolution is not like a smooth, linear, process of design and manufacture, but is full of processes comparable with those engineers would call "tinkering" and "kludging."[9] A common occurrence in evolution is that a trait that is adaptive because serving one function is co-opted for another, and so it is not obvious what *the* function of the trait is in the subsequent system. Co-option and functional multi-tasking are reasons why evolved systems have the heterarchical character of interactions ranging across levels. Generally speaking, to the extent that evolution is "inelegant" and divergent from the designs that would be considered rational and perhaps optimal by a human engineer, there is an obstacle to understanding organic systems through reverse engineering. This is a point made by Kitcher (1988) in relation

[7] For example, glial cells—the very numerous kinds of brain cells that do not generate action potentials—were long thought to be providing metabolic support, but not involved in cognition. This does not appear to be the case, but the challenge of integrating glia into computational theory is immense (Kastanenka et al., 2019).

[8] Lillicrap and Kording (2019) also argue against the comparison between coarse-graining methods in physics and computational explanation in neuroscience.

[9] We should note here that Ballard's representation of software systems as neat and pyramidal is itself an idealization, since large programs like Microsoft Word are themselves the result of years of tinkering and kludging of previous versions of the code.

to Marr's levels, and is reiterated by these biologists more recently:

> deep degeneracy at all levels is an integral part of biology, where machineries[10] are developed through evolution to cope with a multiplicity of functions, and are therefore not necessarily optimized to the problem that we choose to reverse engineer. Viewed in this way, our limitation in reverse engineering a biological system might reflect our misconception of what a design principle in biology is. There are good reasons to believe that this conclusion is generally applicable to reverse engineering in a wide range of biological systems (Marom et al., 2009, p. 3).

Of course, Dennett is aware that the strong assumption of optimality cannot be expected to hold in many cases, but he would advocate for it as a first approximation: the initial prediction is that the evolved system conforms to the expectation based on optimality considerations, and then we look for divergences from this prediction. In this way, reverse engineering retains its heuristic value for biology.

However, we might become less sanguine about the value of this strategy as a heuristic, the more we attend to the worry that cases of conformity to the predictions based on human design considerations are likely to be rare—the first approximation is likely to be just too wide off the mark. On signaling networks in living cells, Moss (2012) points to research findings of everything "cross-talking" to everything else. Such networks are nowhere near the ideal of a hierarchical and near decomposable system. Application of a neat, leveled explanatory framework would only be Procrustean. Both Moss (2012) and Nicholson (2019, p. 115) point to a problem with the wiring diagrams commonly used to represent such networks, based on an analogy with electronic networks, because they lead researchers to underestimate the dynamic nature of these signaling pathways, in comparison with a fixed circuit structure.[11] There is a felt need for better analogies, but perhaps they will not be available for the very reason that human engineered systems—at least when they are intelligible enough to usefully serve as analogies—are too fundamentally different from the evolved ones.

A somewhat controversial view on what is distinctive about natural systems, such that the assumption of near decomposability does not hold, is that they show emergence, meaning that higher level structures impose downward causation on their component parts (Green, 2018). On this view, living systems do have leveled architectures, though radically different from the ones found in artifacts for which the assumption of near decomposability does hold. It is interesting to note that there are new frameworks for engineering, which allow for machines to assemble themselves rather than be constructed according to a transparent, rational plan. It has been argued that some of these artifacts are not modular and near decomposable, and that they may show emergence (see section "Conclusion").

To summarize, my considerations about the difference between living systems and artifacts, boil down to a concern about oversimplification. By making the assumption that living systems such as the nervous system have distinct levels of organization (without downward causation), and using this to justify leveled frameworks in neuroscientific explanation, the density and complexity of brain interactions are most likely being vastly under-estimated. Perhaps this does not matter for a range of predictive and technical purposes, but it does undermine more ambitious claims of level-based theories to be unlocking the riddles of information processing in the brain. Potochnik (2021, p. 24) states the general worry in a compelling way:

> our adherence to the levels concept in the face of the systematic problems plaguing it amounts to a failure to recognize structure we're imposing on the world, to instead mistake this as structure we are reading off the world. Attachment to the concept of levels of organization has, I think, contributed to underestimation of the complexity and variability of our world, including the significance of causal interaction across scales. This has also inhibited our ability to see limitations to our heuristic and to imagine other contrasting heuristics, heuristics that may bear more in common with what our world turns out to actually be like.

The prospect of alternative heuristics is the loaded question. Better notions of levels may yet arise from multi-scale modeling in systems biology. But it could well be that the over-simplifications imposed by artifact analogies and traditional level frameworks are indispensable for making such complex biological systems intelligible to human scientists, given our finite cognitive capacities. In which case, there may be no overall improvement in the heuristics, because any attempts to get closer to the actual complexity of the targets result in a loss of tractability and intelligibility. In which case researchers can, without condemnation, settle for the heuristics that they have, but they should uncouple advocacy of their modest explanatory utility from any stronger claims about brains being computers or organisms being machines.

[10] It is interesting that these scientist use the term "machineries" to refer to biological processes, even when their aim is to draw attention to the limitations of reverse engineering.

[11] "Perhaps the most significant barrier to appreciating the dynamic, heterogeneous aspect of signaling complexes is the lack of a good analogy from our daily experience. This contributes to a second related problem, our inability to depict such interactions diagrammatically. Indeed, the typical "cartoon" of signaling pathways, with their reassuring arrows and limited number of states could be the real villain" (Mayer et al., 2009, p. 6, quoted in Moss, 2012, p. 170).

Conclusion

In this special issue, Bongard and Levin (2021) argue, against Nicholson (2019), that twenty first century machines, such as deep convolutional neural networks (DCNN's), do not have the rigid, modular qualities that, according to Nicholson, make them misleading as models for biological systems. What Bongard and Levin do not consider is that the utility of the analogies is likely to decline once reference is made to self-organizing devices like DCNN's, which do not have the intelligibility of simpler, explicitly designed machines. While the analogy between organisms and machines may become tighter, with the development of machines that are more life-like—that are not modular, and which lack a clear hardware/software division—the motivation for drawing the analogies in the first place may evaporate. For, I have argued in this essay that the payoff of thinking about brains in terms of machine-based comparisons is that it aids explanation by framing the biological object in terms of transparent principles of human-led design. Self-organizing machines lack this attractive transparency. That machines would 1 day become inscrutable was a situation long ago envisaged by one of the first proponents of artificial intelligence and artificial life, John von Neumann:

> At the Hixon Symposium, finding himself taxed by the neurophysiologists ... for not stressing enough the difference between natural and artificial automata, he replied that this distinction would grow weaker over time. Soon, he prophesied, the builders of automata would find themselves as helpless before their creations as we ourselves feel in the presence of complex natural phenomena (Dupuy, 2009, p. 142).

That said, we should not be tempted to conclude that self-organizing, twenty first century machines are absolutely life-like. The problem is that given our relative ignorance about how they work, in comparison with classical machines, we risk also being left in the dark about all the ways they too are *not* like organisms.

Author contributions

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

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ballard, D. (2015). *Brain Computation as Hierarchical Abstraction*. Cambridge, MA: MIT Press.

Barlow, H. (1972). 'Single units and sensation: A neuron doctrine for perceptual psychology?'. *Perception* 1, 371–394. doi: 10.1068/p010371

Bechtel, W., and Bich, L. (2021). 'Grounding cognition: Heterarchical control mechanisms in biology'. *Philos. Trans. R. Soc. B* 376:20190751. doi: 10.1098/rstb.2019.0751

Bechtel, W., and Richardson, R. C. (2010). *Discovering Complexity*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/8328.001.0001

Bongard, J., and Levin, M. (2021). 'Living things are not (20th century) machines: Updating mechanism metaphors in light of the modern science of machine behavior'. *Front. Ecol. Evol.* 9:650726. doi: 10.3389/fevo.2021.650726

Breitenbach, A. (2014). "'Biological purposiveness and analogical reflection," in *Kant's Theory of Biology*, eds I. Goy and E. Watkins (Berlin: Walter De Gruyter). doi: 10.1515/9783110225792.131

Bullock, T. H., Bennett, M. V., Johnston, D., Josephson, R., Marder, E., and Field, R. D. (2005). 'The neuron doctrine, redux. *Science* 310, 791–793. doi: 10.1126/science.1114394

Cao, R. (2014). 'Signaling in the brain: In search of functional units'. *Philos. Sci.* 81, 891–901. doi: 10.1086/677688

Cheung, T. (2006). 'From the organism of a body to the body of an organism: Occurrence and meaning of the word 'oganism' from the seventeenth to the nineteenth centuries'. *Br. J. Hist. Sci.* 39, 319–339. doi: 10.1017/S0007087406007953

Chirimuuta, M. (2021). "'Your brain is like a computer: Function, analogy, simplification," in *Neural Mechanisms: New Challenges in the Philosophy of Neuroscience*, eds F. Calzavarini and M. Viola (Berlin: Springer). doi: 10.1007/978-3-030-54092-0_11

Chirimuuta, M. (under contract). *The Brain Abstracted: Simplification in the History and Philosophy of Neuroscience*. Cambridge, MA: MIT Press.

Cummins, R. (1975). 'Functional analysis'. *J. Philos.* 72, 741–765. doi: 10.2307/2024640

Cummins, R. (2000). "How does it work?" Versus "What are the laws?": Two conceptions of psychological explanation," in *Explanation and Cognition*, eds F. C. Keil and R. A. Wilson (Cambridge, MA: MIT Press).

Dennett, D. C. (1995). "Cognitive science as reverse engineering: Several meanings of "Top-Down" and "Bottom-Up," in *Proceedings of the Ninth International Congress of Logic, Methodology and Philosophy of Science*, eds D. Prawitz, B. Skyrms, and D. Westerståhl (Uppsala), 680–689. doi: 10.1016/S0049-237X(06)80069-8

Dupuy, J.-P. (2009). *On the Origins of Cognitive Science*. Cambridge, MA: MIT Press.

Godfrey-Smith, P. (2016). 'Mind, matter, and metabolism'. *J. Philos.* 113, 481–506. doi: 10.5840/jphil20161131034

Grant, S. (2018). 'Synapse molecular complexity and the plasticity behaviour problem'. *Brain Neurosci. Adv.* 2, 1–7. doi: 10.1177/23982128188 10685

Green, S. (2018). 'Scale dependency and downward causation in biology'. *Philos. Sci.* 85, 998–1011. doi: 10.1086/699758

Gurney, K. N. (2009). 'Reverse engineering the vertebrate brain: Methodological principles for a biologically grounded programme of cognitive modelling'. *Cogn. Comput.* 1, 29–41. doi: 10.1007/s12559-009-9010-2

Haugeland, J. (1978). 'The nature and plausibility of cognitivism'. *Behav. Brain Sci.* 2, 215–226. doi: 10.1017/S0140525X00074148

Hawkins, J., Ahmad, S., and Cui, Y. (2017). 'A theory of how columns in the neocortex enable learning the structure of the world. *Front. Neural Circuits* 11:81. doi: 10.3389/fncir.2017.00081

Illetterati, L. (2014). "'Teleological judgment: Between technique and nature," in *Kant's Theory of Biology*, eds I. Goy and E. Watkins (Berlin: Walter De Gruyter). doi: 10.1515/9783110225792.81

Kant, I. (1790/1952). *The Critique of Judgement*. Oxford: Oxford University Press.

Kastanenka, K. V., Moreno-Bote, R., De Pittà, M., Perea, G., Eraso-Pichot, A., Masgrau, R., et al. (2019). 'A roadmap to integrate astrocytes into systems neuroscience'. *Glia* 68, 5–26. doi: 10.1002/glia.23632

Kitcher, P. (1988). 'Marr's computational theory of vision'. *Philos. Sci.* 55, 1–24. doi: 10.1086/289413

Lillicrap, T. P., and Kording, K. (2019). '*What Does it Mean to Understand a Neural Network?*. Available online at: https://arxiv.org/abs/1907.06374 (accessed July 10, 2020).

Love, B. C. (2021). 'Levels of biological plausibility'. *Philos. Trans. R. Soc. B* 376:20190632. doi: 10.1098/rstb.2019.0632

Marcus, G., and Freeman, J. (2015). "Preface," in *The Future of the Brain*, eds G. Marcus and J. Freeman (Princeton, NJ: Princeton University Press). doi: 10.1515/9781400851935

Marom, S., Meir, R., Braun, E., Gal, A., Kermany, E., and Eytan, D. (2009). 'On the precarious path of reverse neuro-engineering'. *Front. Computat. Neurosci.* 3:5. doi: 10.3389/neuro.10.005.2009

Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman.

Mayer, B., Blinov, M., and Loew, L. (2009). 'Molecular machines or pleiomorphic ensembles: Signaling complexes revisited'. *J. Biol.* 8:81. doi: 10.1186/jbiol185

Moss, L. (2012). 'Is the philosophy of mechanism philosophy enough?'. *Stud. Hist. Philos. Biol. Biomed. Sci.* 43, 164–172. doi: 10.1016/j.shpsc.2011.05.015

Nicholson, D. J. (2019). 'Is the cell really a machine?'. *J. Theor. Biol.* 477, 108–126. doi: 10.1016/j.jtbi.2019.06.002

Nunziante, A. M. (2020). "'Between laws and norms. Genesis of the concept of organism in leibniz and in the early modern western philosophy," in *Natural Born Monads*, eds A. Altobrando and P. Biasetti (Berlin: Walter de Gruyter). doi: 10.1515/9783110604665-002

Potochnik, A. (2021). "'Our world isn't organized into levels," in *Levels of Organization in Biology*, eds D. Brooks, J. DiFrisco, and W. C. Wimsatt (Cambridge, MA: MIT Press).

Simon, H. (1962). 'The Architecture of Complexity'. *Proc. Am. Philos. Soc.* 106, 467–482.

Simon, H. (1969). *The Sciences of the Artificial*. Cambridge, MA: MIT Press.

frontiers | Frontiers in Ecology and Evolution

MINI REVIEW
published: 29 April 2022
doi: 10.3389/fevo.2022.878729

Brains as Computers: Metaphor, Analogy, Theory or Fact?

*Romain Brette**

Sorbonne Université, INSERM, CNRS, Institut de la Vision, Paris, France

Whether electronic, analog or quantum, a computer is a programmable machine. Wilder Penfield held that the brain is literally a computer, because he was a dualist: the mind programs the brain. If this type of dualism is rejected, then identifying the brain to a computer requires defining what a brain "program" might mean and who gets to "program" the brain. If the brain "programs" itself when it learns, then this is a metaphor. If evolution "programs" the brain, then this is a metaphor. Indeed, in the neuroscience literature, the brain-computer is typically not used as an analogy, i.e., as an explicit comparison, but metaphorically, by importing terms from the field of computers into neuroscientific discourse: we assert that brains compute the location of sounds, we wonder how perceptual algorithms are implemented in the brain. Considerable difficulties arise when attempting to give a precise biological description of these terms, which is the sign that we are indeed dealing with a metaphor. Metaphors can be both useful and misleading. The appeal of the brain-computer metaphor is that it promises to bridge physiological and mental domains. But it is misleading because the basis of this promise is that computer terms are themselves imported from the mental domain (calculation, memory, information). In other words, the brain-computer metaphor offers a reductionist view of cognition (all cognition is calculation) rather than a naturalistic theory of cognition, hidden behind a metaphoric blanket.

Keywords: brain-computer metaphor, algorithms, programs, philosophy, metaphors

OPEN ACCESS

Edited by:
Giorgio Matassi,
FRE 3498 Ecologie et Dynamique des Systèmes Anthropisés (EDYSAN), France

Reviewed by:
John Bickle,
Mississippi State University, United States
Alex Gomez-Marin,
Champalimaud Center for the Unknown, Portugal

****Correspondence:***
Romain Brette
romain.brette@inserm.fr

Specialty section:
This article was submitted to Models in Ecology and Evolution, a section of the journal Frontiers in Ecology and Evolution

Received: *18 February 2022*
Accepted: *11 April 2022*
Published: *29 April 2022*

Citation:
Brette R (2022) Brains as Computers: Metaphor, Analogy, Theory or Fact? Front. Ecol. Evol. 10:878729. doi: 10.3389/fevo.2022.878729

WHAT IS A COMPUTER?

It is common to assert that the brain is a sort of computer. It goes without saying that no one believes that people have a hard drive and USB ports. More broadly, a computer is a machine that can be programmed. A program is a set of explicit instructions that fully specify the behavior of the system in advance ("pro-," before; "-gram," write). Computers can be programmed in many different ways: procedural programming (a series of elementary steps, as in a recipe or the C language), logic programming (using logical propositions as in the language Prolog), and so on. There can be such things as "non-conventional" computers, parallel computers, analog computers, quantum computers, and so on, which execute programs in different ways.

"Programmable machine" is both the common usage and the technical usage of "computer." Let us leave aside the concept of a "machine," which would deserve specific treatment (see e.g., Nicholson, 2019; Bongard and Levin, 2021), and allow for an even broader definition: a computer is a programmable thing. Computer science offers no formal definition of computer: it is the concept of program that unifies much of theoretical computer science. In computability theory, a function *f* is said to be computable if there exists a program that can output *f(x)* given *x* as an input. In computability theory, an undecidable problem is a decision problem for which no program gives a correct answer, such as the halting problem. Complexity theory examines the number of steps that a program takes before it stops, and classifies problems with respect to how this number scales with input size. Kolmogorov complexity is the size of the shortest program that produces a given object.

Richards and Lillicrap (2022) rightfully recommend to clarify the exact definition of computer we use, and they offer "some physical machinery that can in theory compute any computable function." Unfortunately, this definition hides the notion of a programmable machine behind the vagueness of the phrase "can in theory." What does it mean that an object *can* do certain things?

Consider a large (say, infinite) pile of electronic components. For any computable function, one "can in theory" assemble the elements into a circuit that computes that function. But this does not make the pile of components a computer. To make it a computer, one would need to add some machinery to build a particular circuit from instructions given by the user. Certainly, the electronic elements "can in theory" compute any computable function, but in the context of computers, what is meant by "can" is that the computer *will* compute the function *if* it is given the adequate instructions, in other words it is a programmable machine.

In the same way, the fact that any logical function can be decomposed into the operations of binary neuron models (McCulloch and Pitts, 1943) does not make the brain a computer, because the brain is not a machine to assemble neurons according to some instructions, as if neurons were construction blocks. Thus, it is fallacious to assert that the brain is literally a computer on the mere basis that formal neural networks can approximate any function (Richards and Lillicrap, 2022), for this would attribute computerness to a disorganized pile of electronic components or to any large enough group of atoms, and this is neither the common usage nor the technical usage in computer science.

A DUALISTIC ENTITY

As pointed out by Bell (1999), the computer is a fundamentally dualistic entity, where some machinery ("hardware") executes instructions ("software") defined by an external agent. It is exactly in this sense that Wilder Penfield, who discovered the cortical homunculi (sensory and motor "maps" of the body on the cortex), claimed that the brain is literally a computer (Penfield, 1975). Penfield was a dualist: he considered that the brain is literally a computer, which gets programmed by the mind.

Although modern neuroscience is deeply influenced by Cartesian dualism, most neuroscientists do not embrace this type of dualism (Cisek, 1999; Mudrik and Maoz, 2015; Brette, 2019). Therefore, it is generally not believed that the brain gets *literally* programmed by some other entity. Perhaps the brain-computer is "programmed by evolution" or "self-programmed," but these are rather vague metaphorical uses. To give some substance to the statement "the brain is a computer," one needs to identify programs in the brain, and a way in which these programs can be changed arbitrarily.

For example, classical connectionism might propose that the program is the set of synaptic weights, and that some process may change these weights. This view, as any attempt to identify a program in the brain, assumes that the brain can be separated into a set of modifiable elements (software) and a fixed set of processes (hardware) that act on those elements, for otherwise the "program" would not unambiguously specify what it does, i.e., would not be a program at all. But synaptic weights are certainly not the only modifiable elements in the brain. This hardware/software distinction is precisely what Bell (1999) opposed because everything in the brain, or in a biological organism, is "soft": "*a computer is an intrinsically dualistic entity, with its physical set-up designed not to interfere with its logical set-up, which executes the computation. In empirical investigation, we find that the brain is not a dualistic entity.*" A living organism does not simply adjust molecular knobs: it continuously produces its own structure, synapses, and everything else (Varela et al., 1974; Kauffman, 1986; Rosen, 2005; Montévil and Mossio, 2015).

Furthermore, to make the case that the brain is a computer, one must demonstrate that there is a way in which the brain's programs can be changed arbitrarily. The problem with this claim is that it implies some form of agency. If not a distinct mind, then who decides to change the program? One might say that the brain is programmed by evolution to achieve some goals, but unless we believe in intelligent design, we know that evolution is not literally a case of programming but rather the natural selection of random structural changes. One might say that the brain "programs itself," but it is not straightforward to give substance to this claim either, beyond the trivial fact that the structure of the brain is plastic. If this plasticity follows some particular rules, then the "programs" that the brain produces are in fact not arbitrary. And indeed, it is not the case that a cat can "self-program" itself into playing chess. Perhaps it might "in theory" be able to play chess, that is, if we allow some fictional observer to rewire the cat's brain in certain ways, but this is not a case self-programming. In the idea that the cat's brain is a computer, there appears to be a confusion of Umwelts (Gomez-Marin, 2019): an observer might be able to "program" a cat's brain in some sense, but the cat itself cannot.

THEORY, ANALOGY, OR METAPHOR?

Therefore, it is not a fact that brains are computers. It might be a certain type of dualist theory, or a fundamentalist connectionist theory, but those theories are at odds with what we know about the biology of brains. However, in most cases, the statement is not taken literally in the neuroscience literature. Is it an analogy or a metaphor? The distinction is that an analogy is explicit while a metaphor is implicit. It might be occasionally stated that the brain is *like* a computer, but a much more common case in the neuroscience literature is that one speaks of sensory *computation*, *algorithms* of decision-making, *hardware* and *software*, *reading* and *writing* the brain (for measuring and stimulating), biological *implementation*, neural *codes*, and so on. These are clear cases of metaphorical writing, borrowing from the lexical field of computers without explicitly comparing the brain to a computer.

Metaphors can be powerful intellectual tools because they transport familiar concepts to an unfamiliar setting, and they have shaped the history of neuroscience (Cobb, 2020). The linguists Lakoff and Johnson (1980) have shown that metaphors pervade our language and shape the concepts with which we think, even though we usually do not notice it ("to shape" in this sentence and "to transport" in the previous one, both applied to concepts). As the authors emphasized: "*What metaphor does is limit what we notice, highlight what we do see, and provide part of the inferential structure that we reason with.*" It is this inferential structure that deserves closer attention. The brain-computer metaphor might be a "semantic debate" (Richards and Lillicrap, 2022), but meaning is actually important. What do we mean when we say that the brain implements algorithms, and is it true?

A DOUBLE METAPHOR

Before we discuss algorithms in the brain, it is useful to reflect on why the brain-computer metaphor is appealing. The brain-computer metaphor seems to offer a natural way to bridge mental and physiological domains. But it is important to realize that it does so precisely because computer words are themselves mental metaphors. In the seventeenth century, a "computer" was a person who did calculations (Hutto et al., 2018). Later on, by analogy, devices built to perform calculations were called computers. We say for example that computers have "memory," but memory is a cognitive ability possessed by persons: it is people who remember, and then we metaphorically say that a computer "memorizes" some information; but when you open some text file, the computer does not literally remember what you wrote. This is why Wittgensteinian philosophers point out that "*taking the brain to be a computer [. . .] is doubly mistaken*" (Smit and Hacker, 2014).

No wonder computers offer a natural way to describe how the brain "implements" cognition: computers were designed with human cognition in mind in the first place. For this reason, there is a sense in which certain persons (but not brains, cats or young children) might literally and trivially be computers: an educated person can execute a series of instructions, for example the integer multiplication algorithm. This trivial sense exists precisely because the computer is modeled on a subset of human cognitive abilities, namely doing calculations. But of course, the relevant scientific question is whether all cognitive activity is of this kind, that is, is a sort of unconscious calculation. In other words, the brain-computer metaphor is a reductionist view of cognition, which claims that all cognitive activity in all animal kingdom (perception, decision, motor control, etc.) is actually composed of elementary cognitive steps, these steps being those displayed by educated humans when they calculate.

At the very least, this claim is not trivially true.

ALGORITHMS OF THE BRAIN

What do we mean when we say that the brain implements algorithms? The textbook definition of algorithm in computer science is: "*a sequence of computational steps that transform the input into the output*" (Cormen et al., 2009). There are different ways to define those steps, but it must be a procedure that is reducible to a finite set of elementary operations applied in a certain order.

What is *not* algorithmic is, for example, the solar system. The motion of planets follows some laws, but it cannot be decomposed into a finite set of operations. These laws constitute a *model* of planet motion, not an algorithm. In the same way, a feedback control system is not in general an algorithm (see e.g., van Gelder's example of Watt's centrifugal governor; van Gelder, 1995). Of course, some algorithms can be feedback control systems, but the converse is not true.

In the same way, a model of brain function is not necessarily an algorithm. Of course, some are. For example, networks of formal binary neurons (McCulloch and Pitts, 1943) are algorithmic. Each "neuron" is defined as a binary function and a feedforward network transforms an input into an output by a composition of such functions. The same applies to deep learning models. Backpropagation is an algorithm too. But the Hodgkin-Huxley model (Hodgkin and Huxley, 1952) is not an algorithm. It is, as the name implies, a model: laws that a number of physical variables obey.

Of course, the Hodgkin-Huxley model can be *simulated* by an algorithm. But the membrane potential is not in reality changed by a sequence of Runge-Kutta steps. More generally, the fact that a relationship between two measurable variables is computable does not imply that the physical system actually implements an algorithm to map one variable to the other. It only means that *someone* can implement the mapping with an algorithm.

Biophysical models of the brain are typically dynamical systems. But dynamical systems are not generically algorithms, and therefore asserting that the brain runs algorithms is a particular commitment that deserves proper justification. To justify it, one needs to identify elementary operations in the brain. For example, the computational view of mind holds that cognition is the manipulation of symbols, that is, the elementary operations are symbolic operations (Pylyshyn, 1980; Shagrir, 2006). This leaves the issue of identifying symbols in the brain, which is generally done through the concept of "neural codes," but this concept is problematic both theoretically and empirically (Brette, 2019). Among other examples, Minsky (1988) attempted to describe cognition in terms of elementary cognitive operations, and Marr (1982) tried to describe vision as a sequence of well-identified signal processing operations, with limited success (Warren, 2012). More generally, it is not so obvious that behavior can be entirely captured by algorithms (Dreyfus, 1978; Roli et al., 2022).

The word "algorithm" is sometimes used in a broader sense, to mean some kind of detailed quantitative description of brain function. But this metaphorical use is confusing: not everything lawful in the world is algorithmic. A quantitative description is a model, not an algorithm, and there are many kinds of model.

COMPUTATION IN THE BRAIN

Perhaps a less misleading term is "computation." The brain might not be a computer, because it is not literally programmable, and it might not literally run algorithms, but it certainly computes: for example, it can transform sound waves captured at the ears into the spatial position of a sound source. But what do we mean by that exactly?

If what we mean is that we are able to locate sounds, look at their expected position and generally behave as a function of source position, then should we not just say that we can perceive the position of sound sources? The word "computation" certainly suggests something more than that. But if so, then this is not a trivial statement and it requires proper justification. Perhaps what is meant is that perception is the result of a series of small operations, that is, by an algorithm, but this is far from obvious.

Perhaps we mean something broader: the brain transforms the acoustic signals into some neural activity that can be identified to source position, and that then leads to appropriate behavior and percepts. But this assumes some form of separability between an encoding and a decoding brain, which can be questioned (Brette, 2019). Or perhaps "computation" is simply meant to designate a transformation from sensory signals to some mental entity that represents source position. The difference between a computation and a mere transformation is then the fact that the output is a representation, not just a value. As Fodor noted, "*there is no computation without representation*" (Fodor, 1981). But then we need to explain what "representation" means in this context, for example that a representation has a truth value (it is correct or not), and how representations relate to brain activity.

Thus, it is not at all obvious in what sense the brain "computes," if it does, and the metaphorical use of the word tends to bury the important questions.

CONCLUSION

Computers are programmable things. Brains are not—at least not literally.

Except in rare Cartesian views where the mind is seen to program the brain (Penfield, 1975), the brain-computer metaphor is indeed a metaphor. Explicit formal comparisons with computers are rare, but brain processes are often described using words borrowed from the lexical field of computers (algorithms, computation, hardware, software, and so on). It is in fact a double metaphor, because computers are themselves metaphorically described with mental terms (e.g., they memorize information). This circular metaphorical relationship explains why the metaphor is (misleadingly) appealing.

The brain-computer metaphor is a source of much confusion in the neuroscience literature, in the same way as the "genetic program" is a source of confusion in genetics (Noble, 2008). "Computer" might be used metaphorically to mean something complicated and useful. But computers run programs: what programs are we referring to? Evolution? The connectome? Neither is actually a program, and it is misleading to suggest they are. "Algorithm" might be used metaphorically to mean "laws" or "model." But this is misleading: "algorithm" suggests elementary operations and codes, which are not found in all models, and certainly not obviously found in brains (Brette, 2019). "Computation" is used metaphorically, but what is meant exactly is generally undisclosed: is it a claim about the algorithmic nature of cognition? about representations? or simply about the fact that behavior is adequate?

Once the meanings of these computer terms are properly disclosed, the scientific debate might begin.

AUTHOR CONTRIBUTIONS

RB wrote the text.

FUNDING

This work was supported by Agence Nationale de la Recherche (Grant Nos. ANR-20-CE30-0025-01 and ANR-21-CE16-0013-02), Programme Investissements d'Avenir IHU FOReSIGHT (Grant No. ANR-18-IAHU-01), and Fondation Pour l'Audition (Grant No. FPA RD-2017-2).

REFERENCES

Bell, A. J. (1999). Levels and loops: the future of artificial intelligence and neuroscience. *Philos. Trans. R. Soc. B Biol. Sci.* 354, 2013–2020. doi: 10.1098/rstb.1999.0540

Bongard, J., and Levin, M. (2021). Living Things Are Not (20th Century) Machines: Updating Mechanism Metaphors in Light of the Modern Science of Machine Behavior. *Front. Ecol. Evol.* 9:650726. doi: 10.3389/fevo.2021.650726

Brette, R. (2019). Is coding a relevant metaphor for the brain? *Behav. Brain Sci.* 42:e215. doi: 10.1017/S0140525X19000049

Cisek, P. (1999). Beyond the computer metaphor: behaviour as interaction. *J. Conscious Stud.* 6, 125–142.

Cobb, M. (2020). *The Idea of the Brain: The Past and Future of Neuroscience*. New York, NY: Basic Books.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms, third edition, third edition*. Cambridge, Mass: The MIT Press.

Dreyfus, H. L. (1978). . *What Computers Can't Do: The Limits of Artificial Intelligence, Revised, Subsequent Édition*. New York, NY: HarperCollins.

Fodor, J. A. (1981). The Mind-Body Problem. *Sci. Am.* 244, 114–123.

Gomez-Marin, A. (2019). A clash of Umwelts: anthropomorphism in behavioral neuroscience. *Behav. Brain Sci.* 42:e229. doi: 10.1017/S0140525X19001237

Hodgkin, A. L., and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* 117, 500–544. doi: 10.1113/jphysiol.1952.sp004764

Hutto, D. D., Myin, E., Peeters, A., and Zahnoun, F. (2018). "The Cognitive Basis of Computation: Putting Computation in Its Place," in *The Routledge Handbook of the Computational Mind*, eds M. Sprevak and M. Colombo (London: Routledge), 272–282. doi: 10.1186/1472-6963-13-111

Kauffman, S. A. (1986). Autocatalytic sets of proteins. *J. Theor. Biol.* 119, 1–24. doi: 10.1016/S0022-5193(86)80047-9

Lakoff, G., and Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY: W. H. Freeman and Company.

McCulloch, W. S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133. doi: 10.1007/BF02478259

Minsky, M. (1988). *Society Of Mind*. New York, NY: Simon & Schuster.

Montévil, M., and Mossio, M. (2015). Biological organisation as closure of constraints. *J. Theor. Biol.* 372, 179–191. doi: 10.1016/j.jtbi.2015.02.029

Mudrik, L., and Maoz, U. (2015). "Me & my brain": exposing neuroscience's closet dualism. *J. Cogn. Neurosci.* 27, 211–221. doi: 10.1162/jocn_a_00723

Nicholson, D. J. (2019). Is the cell really a machine? *J. Theor. Biol.* 477, 108–126. doi: 10.1016/j.jtbi.2019.06.002

Noble, D. (2008). *The Music of Life: Biology Beyond Genes*. Oxford: OUP Oxford.

Penfield, W. (1975). *The Mystery of the Mind*. Princeton: Princeton University Press.

Pylyshyn, Z. W. (1980). Computation and cognition: issues in the foundations of cognitive science. *Behav. Brain Sci.* 3, 111–132. doi: 10.1017/S0140525X00002053

Richards, B. A., and Lillicrap, T. P. (2022). The Brain-Computer Metaphor Debate Is Useless: A Matter of Semantics. *Front. Comput. Sci.* 4:810358. doi: 10.3389/fcomp.2022.810358

Roli, A., Jaeger, J., and Kauffman, S. A. (2022). How Organisms Come to Know the World: Fundamental Limits on Artificial General Intelligence. *Front. Ecol. Evol.* 9:806283. doi: 10.3389/fevo.2021.806283

Rosen, R. (2005). *Life Itself – A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life, New e. Édition*. New York, NY: Columbia University Press.

Shagrir, O. (2006). Why we view the brain as a computer. *Synthese* 153, 393–416. doi: 10.1007/s11229-006-9099-8

Smit, H., and Hacker, P. M. S. (2014). Seven Misconceptions About the Mereological Fallacy: A Compilation for the Perplexed. *Erkenntnis* 79, 1077–1097. doi: 10.1007/s10670-013-9594-5

van Gelder, T. (1995). What Might Cognition Be, If Not Computation? *J. Philos.* 92, 345–381. doi: 10.2307/2941061

Varela, F. G., Maturana, H. R., and Uribe, R. (1974). Autopoiesis: the organization of living systems, its characterization and a model. *Biosystems* 5, 187–196. doi: 10.1016/0303-2647(74)90031-8

Warren, W. H. (2012). Does This Computational Theory Solve the Right Problem? Marr, Gibson, and the Goal of Vision. *Perception* 41, 1053–1060. doi: 10.1068/p7327

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Ecology and Evolution

HYPOTHESIS AND THEORY
published: 16 March 2021
doi: 10.3389/fevo.2021.650726

Living Things Are Not (20th Century) Machines: Updating Mechanism Metaphors in Light of the Modern Science of Machine Behavior

Joshua Bongard[1]† and Michael Levin[2,3]†*

[1] Department of Computer Science, University of Vermont, Burlington, VT, United States, [2] Allen Discovery Center, Tufts University, Medford, MA, United States, [3] Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, United States

One of the most useful metaphors for driving scientific and engineering progress has been that of the "machine." Much controversy exists about the applicability of this concept in the life sciences. Advances in molecular biology have revealed numerous design principles that can be harnessed to understand cells from an engineering perspective, and build novel devices to rationally exploit the laws of chemistry, physics, and computation. At the same time, organicists point to the many unique features of life, especially at larger scales of organization, which have resisted decomposition analysis and artificial implementation. Here, we argue that much of this debate has focused on inessential aspects of machines – classical properties which have been surpassed by advances in modern Machine Behavior and no longer apply. This emerging multidisciplinary field, at the interface of artificial life, machine learning, and synthetic bioengineering, is highlighting the inadequacy of existing definitions. Key terms such as machine, robot, program, software, evolved, designed, etc., need to be revised in light of technological and theoretical advances that have moved past the dated philosophical conceptions that have limited our understanding of both evolved and designed systems. Moving beyond contingent aspects of historical and current machines will enable conceptual tools that embrace inevitable advances in synthetic and hybrid bioengineering and computer science, toward a framework that identifies essential distinctions between fundamental concepts of devices and living agents. Progress in both theory and practical applications requires the establishment of a novel conception of "machines as they could be," based on the profound lessons of biology at all scales. We sketch a perspective that acknowledges the remarkable, unique aspects of life to help re-define key terms, and identify deep, essential features of concepts for a future in which sharp boundaries between evolved and designed systems will not exist.

Keywords: biology, computer science, robot, artificial life, machine learning

OPEN ACCESS

Edited by:
Pedro Martinez,
University of Barcelona, Spain

Reviewed by:
Ricard Sole,
Pompeu Fabra University, Spain
Nora S. Vaage,
Nord University, Norway

****Correspondence:***
Michael Levin
michael.levin@tufts.edu

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Models in Ecology and Evolution, a section of the journal Frontiers in Ecology and Evolution

Received: *07 January 2021*
Accepted: *22 February 2021*
Published: *16 March 2021*

Citation:
Bongard J and Levin M (2021) Living Things Are Not (20th Century) Machines: Updating Mechanism Metaphors in Light of the Modern Science of Machine Behavior. Front. Ecol. Evol. 9:650726. doi: 10.3389/fevo.2021.650726

> "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think."
>
> – Alan Turing, 1950

INTRODUCTION

Living things are amazing – they show resilience, purposeful action, unexpected complexity. They have true "skin in the game" – they actively care about what happens, and can be rewarded or punished by experience. They surprise us at every turn with their ingenuity, their wholism, and their resistance to naïve reductionist approaches to analysis and control. For these reasons, some (Varela and Maturana, 1972; Varela et al., 1974; Rosen, 1985; Nicholson, 2012, 2013, 2014, 2019) have argued against modern cell biology and bioengineering's conceptions of cells as machines (Diaspro, 2004; Davidson, 2012; Kamm and Bashir, 2014). Are living things machines? Defining "life" has proven to be notoriously difficult, and important changes in how we view this basic term have been suggested as a means of spurring progress in the field (Fields and Levin, 2018, 2020; Mariscal and Doolittle, 2020). What is an appropriate definition of "machine," and does it apply to all, some, or no living forms across the tree of life?

Although not unanimously accepted, a powerful view is that all scientific frameworks are metaphors (Honeck and Hoffman, 1980) and the question should be not one of philosophy but of empirical research: does a suitable machine metaphor apply sufficiently to biology to facilitate experimental and conceptual progress? Here we focus attention on common assumptions that have strongly divided organicist and mechanist thinkers with respect to the machine metaphor, and argue that stark classical linguistic and conceptual distinctions are no longer viable or productive. At the risk of making both sides of this debate unhappy, we put our cards on the table as follows. We see life from the organicist perspective (Gurwitsch, 1944; Goodwin, 1977, 1978, 2000; Ho and Fox, 1988; Gilbert and Sarkar, 2000; Solé and Goodwin, 2000; Belousov, 2008). We do not hold reductionist views of the control of life, and one of us (ML) has long argued against the exclusive focus on molecular biology as the only source of order in life (Pezzulo and Levin, 2015, 2016) and the importance of multiple lenses, including a cognitive one, on the problem of biological origins, causation, and biomedical interventions (Manicka and Levin, 2019; Levin, 2020b). However, as often happens, advances in engineering have overtaken philosophical positions, and it is important to re-examine the life-as-machine metaphor with a fair, up-to-date definition of "machine". Our goal here is not to denigrate the remarkable properties of life by equating them with 18th and 19th century notions of machines. Rather than reduce the conception of life to something lesser, we seek to update and elevate the understanding of "machines," given recent advances in artificial life, AI, cybernetics, and evolutionary computation. We believe this will facilitate a better understanding of both – living forms and machines, and is an essential step toward a near future in which functional hybridization will surely erase comfortable, classical boundaries between evolved and engineered complex systems.

Here, we make three basic claims. First, that the notion of "machine" often used to claim that living things are not machines tends to refer to an outdated definition of the term which simply no longer fits. Thus, we have the opportunity (and need) to update the definition of "machine" based on insights from the information, engineering, and life sciences toward a better understanding of the space of possible machines (**Table 1**). We challenge relevant communities to collaborate on a better, more profound definition that makes it clear which aspects fruitfully apply to biological research. Indeed, many other terms such as robot, program, etc. need to be updated in light of recent research trends: these existing concepts simply do not "carve nature at its joints" in the way that seemed obvious in the last century. Second, that progress in the science of machine behavior and in the bioengineering of tightly integrated hybrids between living things and machines breaks down the simplistic dualism of life vs. machine. Instead, we see a continuum of emergence, rational control, and agency that can be instantiated in a myriad of novel implementations, not segregating neatly into categories based on composition (protoplasm vs. silicon) or origin story (evolved vs. designed). Finally, we stress an emerging breakdown not only of distinctions in terminology but of disciplines, suggesting the merging of aspects of information sciences, physics, and biology into a new field whose subject is embodied computation in a very wide range of evolved, designed, and composite media at multiple scales.

WHAT IS MEANT BY "MACHINE"?

To claim that living things are not, or are, machines, it is first necessary to specify what is meant by a "machine" (Turing, 1950; Arbib, 1961; Lucas, 1961; Conrad, 1989; Davidson, 2012; Nicholson, 2012, 2013, 2014, 2019). We view the main aspects of a machine to refer to a device, constructed according to rational principles that enable prediction (to some threshold of accuracy) of their behavior at chosen scales. Machines constrain known laws of physics and computation to achieve specifiable functionality. In addition to this basic description, numerous properties are often assumed and then used to highlight differences from living forms. Let us consider some of these, to understand to what extent they are based on fundamental aspects of what is essential about the concept of machine, not merely contingent aspects due to historical limitations of technological capability. Each of the sections below focuses on one commonly voiced claim regarding the definitions of "machine," which we think is in need of revision in light of advances in the science of machine behavior.

Machines Are Independent: Life Is Interdependent

The Turing machine, a theoretical construct of which all computers are physical instantiations, demonstrated that a clear demarcation exists between a machine and its environment: input and output channels mediate between them. This conception of machines also reaches back further, into the industrial revolution, when mechanical devices formed a new class of matter alongside those of inanimate, animate, and divine phenomena: from the outset, machines were considered as something apart, both from the natural world and from each other. In contrast, living systems are deeply interdependent with one another, simultaneously made and maker. Similarly, the Internet, and now the Internet of Things, is demonstrating that more useful

TABLE 1 | A summary of past differences between machines and living systems and proposed updates that blur the boundaries.

Properties of classical machines that don't apply to life:	Current and future machines are not distinguished from life because:	Proposed new emphasis:
Structure is single-level	Built as multi-scale systems of active, goal-seeking components	Machines composed of parts with self-similar structure and function
Described by a pre-determined list of parts	Protean machines add or subtract components as needed	Machines that make machines of increasing complexity
Machines arise from a design "blueprint"	Self-organizing systems modify their own structure on the fly	Great emphasis on self-controlled allostasis
Tightly constrained operation toward pre-determined functions	Goals are acquired and modified by AI and similar systems	Godel, Turing, deterministic chaos, and other limits apply to predictability and control in machines just as they do in living forms
Highly efficient operation	Noise is exploited, and fallibility of components are expected	Achieving wholistic certainty from uncertain parts.
Function can be interrupted and restarted	Machines modify/improve/complexify their internal structure on the fly	Synergies between useful function and dynamic homeostasis
Behavior is predictable and linear	Perverse instantiation and creativity increasingly result in machines that are not predictable bottom-up	Machines that can perform a desired task in increasingly diverse ways

machines can often be built by composing simpler machines into ever-more complex interdependencies. Modern physical machines are composed of vast numbers of parts manufactured by increasingly interconnected industrial ecologies, and most of the more complex parts have this same property.

Likewise, software systems have very long dependency trees: the hierarchy of support software that must be installed in order for the system in question to run. Software systems are often not considered machines, but rather something that can be executed by a particular class of machine: the Turing machine. However, modern computer science concepts have blurred this distinction between software and machine. A simple example is that of a virtual machine, which is software that simulates hardware different from that running the virtual machine software. This in turn raises the question of whether there is a distinction between simulating and instantiating a machine, but this deep question will be dealt with in forthcoming work.

Moreover, some machines are now becoming part of highly integrated novel systems with living organisms, for sensory augmentation (Sampaio et al., 2001), brain-machine interfaces (Danilov and Tyler, 2005; Shanechi et al., 2014), brain implants to manage epilepsy, paralysis, and other brain states (Shanechi et al., 2014; Alcala-Zermeno et al., 2020), performance augmentation (Suthana et al., 2012; Salvi et al., 2020), and internal physiological homeostatic devices [e.g., increasingly more intelligent devices to manage context-specific, homeostatic delivery of insulin, neurotransmitters, etc (Lee et al., 2019)]. Machines (such as optogenetics interfaces with machine learning components) can even be used to read memories or incept them directly into biological minds (Shen et al., 2019a,b; Vetere et al., 2019), bypassing traditional mechanisms of perception, memory formation, and communication, to access the core of what it means to be a sentient agent. These biohybrid machines require a constellation of particularly dense software and hardware support, maintenance and monitoring, since any cessation of function could injure or mortally endanger a human wearer. In their more exotic implementations, hybridized biological tissue (including brains) with electronics provide a plethora of possible constructs in which obviously alive components are tightly interweaved, in both structure and function, with machine components (Green and Kalaska, 2011; Wilson et al., 2013; Pais-Vieira et al., 2015). The function, cognition, and status of these hybrid systems make clear that no simple dichotomy can be drawn between life and machine.

Machines Are Predictable: Life Is Unpredictable

Intuitively, a useful machine is a reliable one. In contrast, living systems must be noisy and unpredictable: a reliable organism can be easily predated upon; a stationary species can be out-evolved. But, emerging technologies increasingly achieve reliable function by combining uncertain events in novel ways. Examples include quantum computers and machine learning algorithms peppered with stochastic events to ensure learning does not become trapped in partial solutions (Kingma and Ba, 2014). We are also now learning that unpredictability in the long run is often the signature of particularly powerful technologies. Indeed, the inability to predict the "killer app" for a new technology such as a quantum computer or driverless cars is often a signature of particularly disruptive technologies. The utility of surprising machines has historical roots: Gray Walter's physical machines (Walter, 1950) and Braitenberg's hypothetical machines were capable of startlingly complex behavior despite their extreme simplicity (Braitenberg, 1984). Today, robot swarms are often trained to exhibit useful "emergent behavior," although the global behavior of the swarm may not be surprising, the irreducibility of swarm behavior to individual robot actions is a new concept to many roboticists (McLennan-Smith et al., 2020). Finally, the ubiquity of perverse instantiation – automatically trained or evolved robots often instantiate the requested, desired behavior in unexpected ways - in AI has been cited as a potentially useful way of designing machines (Lehman et al., 2020).

Nicholson (2019) defined a machine as having four clear specifications. For 21st century machines, it is becoming increasingly difficult to write down a clear set of specifications for them which spans all the possible ways in which they may change, and be changed by their increasingly complex environments. Instead, it is more useful to think about specifications for the algorithms that then build machines much more complex than the algorithms: canonical examples include the "specification" of the backpropagation of error algorithm that trains deep networks, and the traversal of a search space by genetic algorithms.

Machines Are Designed by Humans: Life Is Evolved

Almost all machines have a human provenance; whereas the very definition of a living system is that it arose from an evolutionary process. Somewhat surprisingly, economic theory provided one of the first intuition pumps for considering the non-human generators of machines: machines arise from the literal hands of human engineers but also the "invisible hand" of a free market; the latter set of pressures in effect "select," without human design or forethought, which technologies proliferate (Beinhocker, 2020). More recently, evolutionary algorithms, a type of machine learning algorithm, have demonstrated that, among other things, jet engines (Yu et al., 2019), metamaterials (Zhang et al., 2020), consumer products (Zhou et al., 2020), robots (Brodbeck et al., 2018; Shah et al., 2020), and synthetic organisms (Kriegman et al., 2020) can be evolved rather than designed: an evolutionary algorithm generates a population of random artifacts, scores them against human-formulated desiderata, and replaces low-scoring individuals with randomly modified copies of the survivors. Indeed, the "middle man" has even been removed in some evolutionary algorithms by searching for novelty rather than selecting for a desired behavior (Lehman and Stanley, 2011). Thus, future agents are likely to have origin stories ranging across a very rich option space of combinations of evolutionary processes and intelligent design by humans and other machines.

The cost of evolving useful machines, rather than designing them by hand, is that that they are often inefficient. Like organisms, evolved machines inherently include many sub-functions exapted from sub-functions in their ancestral machines, or mutations that copy and differentiate sub-functions leads to several modules with overlapping functions. Nicholson (2019)'s third necessary feature of machines is that they are efficient: again, 21st century machines increasingly lack this property. The increased use of evolutionary dynamics by engineers, and the ability of both kinds of processes to give rise to highly adapted, complex systems makes it impossible to use evolved vs. designed as a clear demarcation between two classes of beings.

An especially powerful blow to the conceit that machines are the direct result of human ingenuity are machines that make machines. Mass production provided the first example of a machine — a factory – that could produce other machines. John von Neumann postulated theoretical machines that could make perfect copies of themselves, which in turn make copies of themselves, indefinitely, assuming a constant supply of raw building materials (von Neumann and Burks, 1966). Theory has been partly grounded in practice by rapid prototyping machines that print and assemble almost all of their own parts (Jones et al., 2011). Similarly, many are comfortable with the idea that the Internet, a type of machine, helps "birth" new social network applications. Those applications in turn connect experts together in new ways such that they midwife the arrival of brand new kinds of hardware and software. Indeed, most new technologies result from complex admixtures of human and machine effort in which economic and algorithmic evolutionary pressures are brought to bear. Indeed, one defensible metric of technological progress is the growing number of intermediate machine design/optimization layers sandwiched between human ingenuity and deployment of a new technology.

Life Is Hierarchical and Self-Similar: Machines Are Linearly Modular

Living systems exhibit similar structure and function at many different levels of organization. As one example of self-similar structure, at small scales, branching structures are not just self-similar but even fractal. Another example is the interdependence between hierarchical structure and function in the brain (Sporns et al., 2000). Even more important is self-similar function, in the sense of multi-scale competency, allostasis, or homeostasis (Vernon et al., 2015; Schulkin and Sterling, 2019): organelles, cells, organisms, and possibly species evolved adaptive mechanisms to recover when drawn away from agreeable environmental conditions or even placed in novel circumstances. Machines are typically assumed to be hierarchical and modular for sound engineering reasons, but self-similarity in machines is less obvious. Although fractality is currently under investigation in software (Semenov, 2020), circuit design (Chen et al., 2017), and metamaterials (De Nicola et al., 2020), it is conspicuously absent from other classes of machines. Thus, unlike the other features considered in this section, self-similarity remains a feature that, for now, does tend to distinguish living systems from machines. It is important to note, however, that this is not fundamental – there is no deep reason that prevents engineered artifacts from exploiting the deep, multi-scale organization of living organisms to improve problem-solving and robustness. Although many current machines are highly modular and efficient by design, machines produced by other machines increasingly exhibit differing amounts and types of modularity. Indeed artificially evolved neural networks (Clune et al., 2013) and robots (Bernatskiy and Bongard, 2017) often lack modularity unless it is directly selected for, and many exhibit inefficiencies caused by evolutionarily duplicated and differentiated sub-structures and sub-functions (Calabretta et al., 2000).

Autonomy at many scales is especially important with respect to function, not only structure (Pezzulo and Levin, 2016; Fields and Levin, 2017, 2020). Biological systems are holarchies in which each subsystem is competent in achieving specific goals (in the cybernetic, allostatic sense) despite changing local circumstances (Pezzulo and Levin, 2016). For example, a swarm of tadpoles organizes its swimming in a circular pattern to ensure efficient flow of nutrients past their gills. At the same time, individual tadpoles perform goal-directed behaviors and compete with each other, while their craniofacial organs re-arrange toward a specific target morphology of a frog [able to pursue this anatomical goal regardless of their starting configuration (Vandenberg et al., 2012)], their tissues compete for informational and nutritional resources (Gawne et al., 2020), and their individual cells maintain metabolic and homeostatic and transcriptional goal states. Such nested architecture of competing and cooperating units achieves unprecedented levels of robustness, plasticity, and problem-solving in novel circumstances (Levin, 2019; Levin, 2020a). It is also likely responsible for the remarkable evolvability of living forms, because such multi-scale competency flattens the fitness landscape: mutations have fewer deleterious effects if some of the changes they induce can be compensated by various subsystems, allowing their negative effects to be buffered while the positive effects accumulate. At present, this is a real difference between how we engineer machines and how living things are constructed; for now, defections of parts from the goals of the whole system (robot "cancer") are rare, but this will not be the case for long. We expect near-future work to give rise to machines built on the principles of multi-scale competency in a fluid "society" of components that communicate, trade, cooperate, compete, and barter information and energy resources as do living components of an organism (Gawne et al., 2020).

Life Is Capable of Intelligence (And Free Will, Subjectivity, Consciousness, Agency, and Metacognition): Machines Are Not; Indeed, They Never Will

Nowhere does the specter of Cartesian dualism loom more prominently than in the debates about whether current machines possess any of the cognitive and affective features usually associated with higher animals, such as intelligence, agency, self-awareness, consciousness, metacognition, subjectivity, and so on (Cruse and Schilling, 2013). Indeed, the most intense debates focus on whether machines will ever be able to attain one or more of these internal states. As many have pointed out, the stronger the claim that higher cognition and subjectivity is only accessible to living systems, the stronger the evidence required to prove that living systems possess them. It is still strongly debated what aspects of the body organization are required for these capacities, or even whether such phenomena exist at all (Lyon, 2006; Bronfman et al., 2016; Dennett, 2017). Until such time as firm definitions of these terms is arrived at, claiming them as a point of demarcation between machines and life is an ill-defined exercise. Moreover, it is now clear that composite, hybrid creatures can be bioengineered with any desired combination of living cells (or whole brains) and real-time optical-electrical interfaces to machine-learning architectures (Grosenick et al., 2015; Newman et al., 2015; Pashaie et al., 2015; Roy et al., 2017). Because the living tissue (which houses the symbol grounding and true "understanding") closely *interacts* with the machine learning components, forming a single integrated system, such chimeras reveal that there is no principled way to draw a crisp line between systems that have true subjectivity and those that are mere engineered systems.

Machines are increasingly occupying new spaces on the scale of *persuadability*, which ranges from low-level, physical control that has to be applied to change the function of a mechanical clock, to the use of experiences (positive or negative reinforcement), signals, messages, and arguments that one can use with agents of increasing cognitive sophistication. One way to formalize this distinction is through the relative amount of energy or effort used in an intervention compared to the change in the system's behavior. Messages, unlike physical pushes, require relatively low energy input because they count on the receiving system to do a lot of the hard work. If one wants a 200 kg block of aluminum to move from point A to point B, one has to push it. If one wants a 200 kg robot to make the same journey, it may be sufficient to provide only a simple signal; and if one is dealing with a human or complex AI, one could even implement the move to occur in the future, in some specific context, by providing a rational reason to do it (via a low-energy message channel (Hoffmeyer, 2000; Pattee, 2001). Modern autonomous machines require increasingly low-energy interventions to produce useful work – a trend begun decades ago by developments in cybernetics. Indeed, in their increasing large-scale lability in the face of very subtle signals, they may get closer to the edge of chaos that is so prevalent in biology (Hiett, 1999; Kauffman and Johnsen, 1991; Mora and Bialek, 2011).

Machines Can Be Studied in a Reductionist Framework: Life Cannot

Until very recently, the very fact that machines could be rapidly disassembled into their component parts, repaired or improved, and then reassembled, was one of their primary advantages over living machines such as domesticated animals or human slaves. This modularity and hierarchy continues today in our most complex technologies, like state-of-the-art computer chips, which contain billions of transistors. Progress in circuit design now requires reaching into the quantum realm (Preskill, 2018), or enlisting DNA to store and transmit information (Chatterjee et al., 2017). Reductionist approaches in Artificial Intelligence are rapidly losing explanatory power as AI systems assume greater complexity. Considering the weight of a particular synaptic connection or a local neural cluster in a deep neural network provides little understanding of the machine's behavior as a whole. Making progress in these domains may incur a cost of not being able to guarantee how local behavior will resolve into global behavior, like computation speed (a feature related to the predictability issues discussed above). Instead, AI methods may have to be enlisted to design such circuits. Ironically, the AI methods and their products, like neural networks, are both extremely resistant to reductionist analysis. As just one example, although the most common form of training neural networks, the backpropagation of error, is a simple mathematical technique, one of the co-founders of this method and other AI "insiders" have admitted to being baffled at its surprising effectiveness (Sejnowski, 2020). As for AI's objects – neural networks – the very nature of their immense interconnectivity frustrates most attempts to summarize their global behavior by only referring to the individual behavior of their edge weights. Indeed, the fact that neural networks are modeled on biological nervous system principles makes it unsurprising that they would exhibit many biological features, including that of resistance to reductionist analysis. Many machines, especially swarms, exhibit behavior that requires the same techniques used to study cognition in biological systems (Beer, 2004, 2014, 2015; Swain et al., 2012; Pavlic and Pratt, 2013; Nitsch and Popp, 2014; Beer and Williams, 2015; Slavkov et al., 2018; Valentini et al., 2018), and even relatively straightforward machines are surprisingly resistant to analysis using today's analytical tools (Jonas and Kording, 2016).

Today's and future autonomous machines, like living things, will be subject to deterministic chaos (amplification of very small differences in initial conditions), inputs from their environment that radically affect downstream responses, highly complex interactions of a myriad diverse internal parts, and perhaps even quantum uncertainty (Thubagere et al., 2017). For the most sophisticated agents, a high level of analysis (in terms of motivations, beliefs, memories, valences, and goals) may be far more effective than bottom-up prediction approach – much as occurs in biology (Marr, 1982; Pezzulo and Levin, 2015, 2016).

If reductionist analysis is impossible for current and future machines, what remains? A consortium of social scientists, computer scientists and ethologists recently called for the creation of a new field, "machine behavior," in which the best explanations of machines, and predictions of their likely behavior, are a combination of wholistic methods drawn from ethology, the social sciences, and cognitive science (Rahwan et al., 2019). As just one example, most modern deep learning analytic methods attempt to discover pathological holistic behavior in neural networks, such as bias. Then, these methods attempt to discover the likely root cause of that behavior and rectify it, such as de-biasing biased training data sets (Bolukbasi et al., 2016). Indeed in many cases, the most effective explanations of animal and human behavior stop far short of detailed neurological, chemical or small-scale physical phenomena (Noble, 2012). This call for wholistic thinking is partly intellectual and partly pragmatic: we require compact, falsifiable and predictive claims about how autonomous machines will act in the world, in close proximity to humans. Such claims provide a firm foundation for new knowledge, but also for new legislation, regulation, and social norms. Finally, the deeply social and, increasingly, biological components of modern machines further complicate reductionist thinking: extrapolating what a million people will do with a million plows, given knowledge of a plow, is tractable. Predicting what 3.8 billion people will do with 3.8 billion social media accounts[1], or an equivalent number of brain-computer interactive devices, is not.

Nicholson (Nicholson, 2019) concluded his list of three necessary features for machines – specificity, constraint, and efficiency – with a fourth and final feature: non-continuity. By this, he meant that machines could be halted, disassembled, understood, repaired, and reassembled. As with the first three features, 21st century machines are increasingly resistant to reductionist manipulations as well as reductionist explanations (Guidotti et al., 2019; Rudin, 2019). Put differently, modern technologies only achieve utility when they are emplaced appropriately into the technosphere; it is difficult or impossible to describe their function independently of it.

Life Is Embodied: AIs Are Not

Above, we have considered increasingly untenable distinctions between machines and living systems. Another commonly voiced distinction inherited from Cartesian dualism, but one which is also rapidly deteriorating in the face of advances in technology, is that between embodied creatures and pure (software-based) AI. The staying power of this distinction is mostly due to its seeming intuitive nature: a living being (or robot) acts directly on the world, and is affected by it; "AI" are programs that run inside a computer and thus only impact the world indirectly. The sharp separation between AIs, whose essential nature is an algorithm (which can be run on many different kinds of hardware) seems categorically different than a living being which is defined by its particulars, in both mind and body. It is curious that a discipline only 70 years old should be so deeply cleaved along fault lines established at the outset of Western thought, millennia ago. Much ink has been spilled on this subject that we will not attempt to summarize here; instead, we will highlight a few thrusts within both disciplines that unintentionally or intentionally attempt to close this gap.

"Embodied AI" has come to be associated with efforts to run deep learning algorithms on autonomous robots (Savva et al., 2019). However, these methods can be seen as deepening rather than narrowing the brain/body distinction: In these approaches, the robot's form is usually a fixed shell, previously designed by human engineers, controlled by the machine learning algorithm. In contrast, there is a small but growing literature on embodying intelligence directly into the body of the robot (Nakajima et al., 2015), and in machine learning methods that evolve robot bodies to enhance this and other forms of intelligence (Powers et al., 2020). A small but growing literature on robots capable of self-modeling also blurs the distinction between embodied robots and non-embodied AI methods. Attempts here focus on enabling a robot to model its own body (Bongard et al., 2006; Kwiatkowski and Lipson, 2019), and model unexpected changes to that body such as damage, using AI methods. In such systems, morphological change is occurring alongside mental changes, such as improved understanding of the robot's current internal and external states (Kwiatkowski and Lipson, 2019). Likewise, an important distinction for the biosciences is between disciplines like zoology, which focus on very specific examples of life, and the study of deep principles of biological regulation ["life as it could

[1] The current estimate of people with social media accounts as of January 2020 (statista.com;bit.ly/2KSdA9U).

be," (Langton, 1995; Walker and Davies, 2013)] which, like AI software, can be implemented in a wide range of media.

Machines Have Clear Hardware/Software Distinctions: Life Does Not

One of the most enduring technological metaphors applied to organisms is that of DNA as software and cells as hardware. The metaphor sometimes considers transcription and translation as the interface between the two. In this guise, transcription and translation serve as the biological equivalent of finite automata, which translate code into physical changes imposed on the world. Biological nervous systems acquire a similar metaphor by extension, but here software is often considered to be electrical activity in the brain. Software, as the name implies, is usually restricted to "fluid" systems: chemical, electrical, or sub-atomic dynamics. Hardware is instead usually applied to macroscale, Newtonian, mechanical objects such as switches and relays in artificial systems, and physiology in living systems. Several advances in neuroscience and regenerative biology challenge the claim that biology never exploits the software/hardware distinction. For example, it has been argued that changes in blood flow in the brain can convey information (Moore and Cao, 2008), as does the function of astrocytes (Santello et al., 2019) and neurotransmitters (Ma et al., 2016). The non-electrical components of these structures and mechanisms complicate extending the software metaphor to encompass them. The hardware/software distinction is also blurring in technological systems: increasingly specialized hardware is being developed to support deep learning-specific algorithms (Haensch et al., 2018), and the physics of robot movement can be considered to be performing computation (Nakajima et al., 2015). DNA computing further complicates the hardware/software distinction: In one recent application (Chatterjee et al., 2017), DNA fragments simultaneously house the "software" of a given species yet also serve as logic gates and signal transmission lines, the atomic building blocks of computer hardware. Robots built from DNA (Thubagere et al., 2017) reduce the distinction yet further (Thubagere et al., 2017). Moreover, recent work on bioelectric control of regenerative setpoints showed that planarian flatworms contain voltage patterns (in non-neural cells) that are not a readout of current anatomy, but are a re-writable, latent pattern memory that will guide regenerative anatomy if the animal gets injured in the future (Levin et al., 2018). These patterns can now be re-written, analogous to false memory inception in the brain (Ramirez et al., 2013; Liu et al., 2014), resulting in worms that permanently generate 2-headed forms despite their completely wild-type genetic sequence (Durant et al., 2017). This demonstrates a sharp distinction between the machine that builds the body (cellular networks) and the data (stable patterns of bioelectric state) that these collective agents use to decide what to build. The data can be edited in real time, without touching the genome (hardware specification).

Most recently, the authors' work on computer-designed organisms (Kriegman et al., 2020) calls this distinction into question from another direction. An evolutionary algorithm was tasked with finding an appropriate shape and tissue distribution for simulated cell clusters that yielded the fastest self-motile clusters in a virtual environment. A cell-based construction kit was made available to the algorithm, but it was composed of just two building units: *Xenopus laevis* epithelial and cardiac muscle cells. The fastest-moving designs were built by microsurgery using physical cells harvested from *X. laevis* blastulae. The resulting organism's fast movement, with anatomical structure and behavior entirely different from that of normal frog larvae, was thus purely a function of its evolved, novel shape and tissue distribution, not neural control or genomic information. Such an intervention "reprograms" the wild type organism by forcing it into a novel, stable, bioelectric/morphological/behavioral state, all without altering the DNA "software." This inverts the normal conception of programming a machine by altering its software but not its hardware.

IMPROVING DEFINITIONS

Given the increasingly unsupportable distinctions between machines and life discussed above, we suggest that updated definitions of machine, robot, program, software, and hardware are in order. The very fact that many of these systems are converging makes delineating them from one another an almost paradoxical enterprise. Our goal is not to etch in stone precise new definitions, but rather to provide an update and starting point for discussion of terms that often are used without examination of their limitations. We emphasize aspects that we hope summarize important emerging structure and commonalities across these concepts. Wrestling with these concepts helps identify previously unasked research questions and unify research programs that previously were treated as distinct with respect to funding bodies, educational programs, and academic and industrial research environments.

Machine

Any system that magnifies and partly or completely automates an agent's ability to effect change on the world. The system should be composed of parts several steps removed from raw materials and should be the result of a rational, or evolutionary (or both), design process. Importantly, a machine uses rationally discoverable principles of physics and computation, at whatever level (from molecular to cognitive), to achieve specific functions and is controllable by interventions either at the physical level or at the level of inputs, stimuli, or persuasion via messages that take advantage of its computational structure. The definition would include domesticated plants and animals (systems with rationally modified structure and behavior), and synthetic organisms. Machines often have exhibit information dynamics that enhance an agent's ability to effect change on the world. The agent may be the entity who constructed the machine, or a third party. Similarly, the agent need not be self-aware or even sentient.

We propose that physicality is not a requirement. Physicality too easily becomes a seemingly obvious, Cartesian border between one class of phenomena and another. Of more interest are machines in which small-scale physical phenomena, such as quantum and electrodynamic forces in biological cells or microscale robots, influence macroscale behavior, such as whole-body motion or swarm intelligence. By removing the physicality requirement, a machine may be a machine learning algorithm that generates better machine learning algorithms or designs robots or synthetic organisms.

Robot

A machine capable of physical actions which have direct impacts on the world, and which can sense the repercussions of those actions, and is partly or completely independent of human action and intent. This definition is related to embodiment and situatedness, two previous pillars supporting the definition of robot (Pfeifer and Bongard, 2006). Crucially, the property of being a robot is not a binary one, but rather a spectrum – a continuum (independent of origin story or material implementation). The determinant of where a given system lands on the continuum is the degree of autonomous control evidenced by the system (Rosen, 1985; Bertschinger et al., 2008). A closely related continuum reflects the degree of *persuadability* of the system (Dennett, 1987). On one end of the continuum are highly mechanical systems that can only be controlled by direct physical intervention – micromanagement of outcome by "rewiring". In the middle are systems that can be stimulated to change their activity – they can be sent signals, or motivated via reward or punishment experiences based on which they can make immediate decisions. At the far end are systems in which an effective means of communication and control is to alter the goals that drive their longer-term behavioral policies – they can be persuaded by informational messages encoding *reasons*, based on which they will change their goals. The important variables here are the causal closure of the system in its behavior (Rosen, 1974; Montevil and Mossio, 2015), and the amount of energy and intervention effort that need to be applied to get the system to make large changes in its function (the smaller the force needed to affect the system, the more sophisticated the robot). A continuous measure of the level of roboticism is required, to handle the growing class of hybrids of biological and mechano-electronic devices. For example, smart prosthetics, which are mostly under human control via muscle activation or thought processes, are less robotic than an autonomous car.

Program

A program is typically conceived as an abstract procedure that is multiply realizable: different physical systems can be found or constructed that execute the program. We see no need to alter this definition, except to state that execution need not be restricted to electrical activity in a computer chip or nervous tissue; chemical (Gromski et al., 2020) and mechanical (Silva et al., 2014) processes may support computation as well. However, a couple of aspects are important for discussing programs in biology. First, that programs do not need to be written by humans, or be a linear one-step-at-a-time procedure – the kinds of programs that (rightly) cause many to say that living things do not follow programs. The set of possible programs is much broader than that, and subsumes distributed, stochastic, evolved strategies such as carried out by nervous systems and non-neural cellular collectives. Indeed the question of whether something is a program or not is relative to a scale of biological organization. For example, genetic sequence is absolutely not a program with respect to anatomical shape, but it is a program with respect to protein sequence.

Software/Hardware

The common names for this technological pairing hint that material properties are what distinguish software from hardware; one can contrast the fluid flow of electrons through circuitry or photons through photonic circuits (Thomson et al., 2016) against the rigidity of metal boxes, vacuum tubes, and transistors. However, another, operational interpretation of their etymology is possible: it is harder to change hardware than software, but examples abound of both, radical structural change (Birnbaum and Alvarado, 2008; Levin, 2020a) and learning/plasticity at the dynamical system level that does not require rewiring (Biswas et al., 2021). Programmable matter (Hawkes et al., 2010) and shape changing soft robots (Shah et al., 2020) are but two technological disciplines investigating physically fluid technologies. In one study the assumption that changing hardware is hard was fully inverted: it was shown that a soft robot may recover from unexpected physical injury faster if it contorts its body into a new shape (a hardware change) rather than learning a compensating gait [a software change; (Kriegman et al., 2019)]. These distortions and inversions of the hardware/software distinction suggest that a binary distinction may not be useful at all when investigating biological adaptation or creating intelligent machines. However, a continuous variant may be useful in the biosciences as follows: a living system is software reprogrammable to the extent that stimuli (signals, experiences) can be used to alter its behavior and functionality, as opposed to needing physical rewiring (e.g., genome-editing, cellular transplantation, surgical interventions, etc.).

We suggest that the low-hanging fruit of specifiable, constrained, efficient, and fully predictable 20th century machines have now been picked. We as a society, and researchers in several fields, can (and must) now erase artificial boundaries to create machines that are more like the structures and processes that life exploits so successfully.

AN EMERGING FIELD: RE-DRAWING THE BOUNDARIES

> "Computer science is no more about computers than astronomy is about telescopes"
>
> – Edsger Dijkstra

The differences that have been cited between living beings and machines are generally ones that can (and will) be overcome by incremental progress. And even if one holds out for some essential ingredient that, in principle, technology cannot copy, there is the issue of hybridization. Biological brains readily incorporate novel sensory-motor (Bach-y-Rita, 1967; Sampaio et al., 2001; Ptito et al., 2005; Froese et al., 2012; Chamola et al., 2020) and information-processing (Clark and Chalmers, 1998) functions provided by embedded electronic interfaces or machine-learning components that provide smart, closed-loop reward neurotransmitter levels (Bozorgzadeh et al., 2016) or electrical activity which can modulate cognition. Even if "true" preferences, motivations, goal-directedness, symbol grounding, and understanding are somehow only possible in biological media, we now know that hybrid functional systems can be constructed that are part living tissue and part (perhaps smart) electronics (Reger et al., 2000; DeMarse and Dockendorf, 2005; Hamann et al., 2015; von Mammen et al., 2016; Ando and Kanzaki, 2020), presumably conferring all of those features onto the system. No principled limits to functionalization between living systems (at any level of organization) and inorganic machinery are known; even if such limits exist, these ineffable components of living things will still tightly interact with engineered components through the interface of other biological aspects of cells and tissues that are already known to be closely interoperable with inorganic machine parts. Thus, we visualize a smooth, multi-axis continuum of beings being made of some percentage of parts that are uncontroversially biological and the remaining percentage of parts that are obviously machines (**Figure 1**). They are tightly integrated in a way that makes the whole system difficult to categorize, in the same way that molecular machines (e.g., ATPase motors or folding-programmed DNA strands) work together to make living beings that implement much more flexible, high-order behavior.

The near future will also surely contain systems in which biological and artificial parts and processes are intermixed across many levels of organization, and many orders of spatial and temporal scales. This could include a swarm composed of robots and organisms, and in which this admixture gradually changes over time to respond to slow time scale evolutionary pressures: the biological units reproduce and evolve, and the mechanical units self-replicate and evolve. Each individual in the swarm may itself be a cyborg capable of dynamically reconfiguring its biological and artificial components, while each of its cells may include more or less genetic manipulation. Where in such a system could a binary dividing line between life and artifice be placed?

Working as coherent wholes, such constructs make highly implausible a view of strict life/machine dualism, in the same way that the problem of explaining interaction vexed Descartes' dualism between body and mind. Thus, the hard work of the coming decades will be to identify what, if any, are *essential* differences – are there fundamentally different natural kinds, or major transitions, in the continuum of fused biological and technological systems? At stake is a conceptual framework to guide basic research and applied engineering in the coming decades, which is essential given the exponential rate of progress in capabilities of altering and hybridizing the products of biology and computer engineering.

Familiar boundaries between disciplines may be more a relic of the history of science than optimal ways to organize our knowledge of reality. One possibility is that biology and computer science are both studying the same remarkable processes, just operating in different media. We suggest that the material implementation and the back-story of a given system are not sufficient information to reliably place it into a category of machine vs. living being, and indeed that those categories may not be discrete bins but rather positions in a multidimensional but continuous space. By asking hard questions about the utility of terminology whose distinct boundaries were calcified centuries ago, a number of advantages will be gained. The obvious trajectory of today's technology will result in the presence of novel, composite creatures that in prior ages could be safely treated as fun sci-fi that didn't have to be dealt with seriously. Updating our definitions and clearly articulating the essential differences between diverse types of systems is especially essential given the aspects of bioengineering and machine learning advances that cannot yet be foreseen.

THE INTERDISCIPLINARY BENEFITS OF A NEW SCIENCE OF MACHINES

The biosciences have much to gain from a more nuanced, non-binary division between life and machine, and the emergence of the field of machine behavior. First, the fact that modern machines are multi-scale, surprising systems that are often as hard to predict and control as living systems (Man and Damasio, 2019; Rahwan et al., 2019) drives improvement in strategies for reverse-engineering, modeling, and multi-level analysis. This is exactly what is needed to break through complexity barriers facing regenerative medicine and developmental biology (Levin, 2020a). For example, solving the inverse problem in biomedical settings (what molecular-level features can be tweaked in order to achieve large-scale outcomes, such as forming an entire human hand via manipulation of gene and pathway activity in single cells) (Lobo et al., 2014; Pezzulo and Levin, 2016) will likely be advanced by the development of engineering approaches to harness noise, unpredictability, and top-down programming of goal-directed multiscale systems.

Second, grappling with issues of control, programmability, agency, and autonomy helps biologists identify and refine essential features of these concepts, freed from the frozen accidents of evolution and the history of biology, where contingent categories (e.g., "consisting of protoplasm") offered distinctions that were easy to use in every-day life but misleading for a deeper scientific understanding. Asking how one can implement intrinsic motivation (Oudeyer and Kaplan, 2007), optimal control (Klyubin et al., 2005), and the ability to pursue and set goals in synthetic constructs (Kamm et al., 2018) will help reveal which aspects of living forms are the wellspring of these capacities and which are contingent details that do not matter.

Third, the engineering and information sciences offer many conceptual tools that should be tested empirically for their utility in driving novel work in basic biology, biomedicine, and synthetic bioengineering. Modular decomposition, software-level reprogrammability, embodied and collective intelligence (Sole et al., 2016), morphological computation (Fuchslin et al., 2013; Corucci et al., 2015), codes and encodings (Barbieri, 1998, 2018, 2019; Levin and Martyniuk, 2018), and much more. Finally, an inclusive, continuous view of life and machines frees the creative

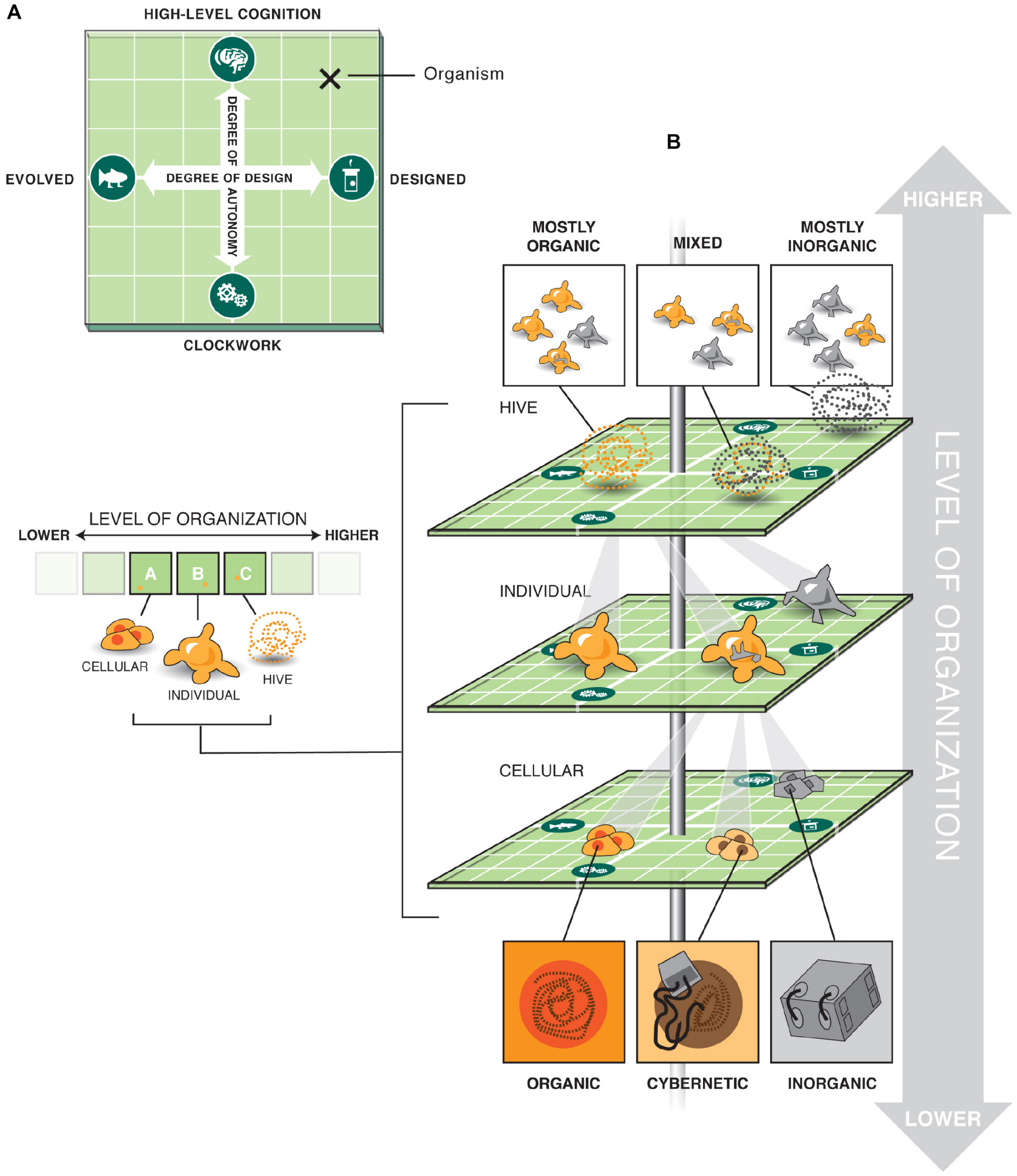


FIGURE 1 | Multi-scale option space for possible living machines. **(A)** Two orthogonal axes define important aspects of any complex system: the degree of design vs. evolution that created it, and the degree of amount of autonomy it is able to implement. We suggest that both of these principal components are not binary categories (such as evolved vs. designed, mechanical vs. autonomous/cognitive) but rather continuous. Together, they form a 2-dimensional option space within which a great variety of possible agents can be placed. **(B)** Importantly, such an option space exists at each level of organization (for example, the familiar biological nested scales of cells, individual organisms, and hives/swarms), and each level comprising a complex agent could occupy a different position in the option space – the levels can be independent with respect to how much evolution, design, and cognition they involve. For example, a given system could be in one corner of the option space at the lowest level (e.g., contain cells that include highly predictable synthetic circuits), but be evolved and intelligent at the level of the individual, and at the same time be part of a swarm containing a mix of designed and evolved agents made up of different elements elsewhere on the option space in **(A)**.

capacity of bioengineers, providing a much richer option space for the creation of novel biological systems via guided self-assembly (Kamm and Bashir, 2014; Kamm et al., 2018). Advances in this field even help address controversies within the biological sciences, such as whether behavior and intelligence are terms that can apply to plants (Applewhite, 1975; Trewavas, 2009; Garzon and Keijzer, 2011; Cvrckova et al., 2016; Calvo et al., 2017).

Likewise, the breaking down of artificial boundaries between the life and engineering sciences has many advantages for computer science and robotics. The first is bioinspiration. Since its cybernetic beginnings, researchers in Artificial Intelligence and robotics have always looked to biological forms and functions for how best to build adaptive and/or intelligent machines. Notable recent successes include convolutional neural networks, the primary engine of the AI revolution, which are inspired by the hierarchical arrangement of receptive fields in the primary visual cortex (Krizhevsky et al., 2017); deep reinforcement learning, the primary method of training autonomous cars and drones, inspired by behaviorism writ large (Mnih et al., 2015); and evolutionary algorithms, capable of producing a diverse set of robots (Bongard, 2013) or algorithms (Schmidt and Lipson, 2009) for a given problem. However, bioinspiration in technology fields is often *ad hoc* and thus successes are intermittent. What is lacking is a systematic method for distilling the wealth of biological knowledge down into useful machine blueprints and algorithm recipes, while filtering out proximate mechanisms that are overly reliant on the natural materials that nature had at hand. The products of research in biology (e.g., scientific papers and models) are often brimming with molecular detail such as specific gene names, and it is an important task for biologists to be able to abstract from inessential details of one specific organism and export the fundamental principles of each capability in such a way that human (or AI-based) engineers can exploit those principles in other media (Slusarczyk et al., 2012; Bacchus et al., 2013; Garcia and Trinh, 2019). Design of resilient, adaptive, autonomous robotics will benefit greatly from importing deep ideas discovered in the principles at work in the biological software that exploits noise, competition, cooperation, goal-directedness, and multi-scale competency.

Second, there is much opportunity for better integration across these fields, both in terms of the technology and the relevant ethics (Levin et al., 2020; Lewis, 2020). Consider the creative collective intelligence that will be embodied by the forthcoming integrated combination of human scientists, *in silico* evolution in virtual worlds, and automated construction of living bodies (Kamm and Bashir, 2014; Kamm et al., 2018; Kriegman et al., 2020; Levin et al., 2020), working together in a closed loop system as a discovery engine for the laws of emergent form and function. All biological and artificial materials and machines strike careful but different balances between many competing performance requirements. By drawing on advances in chemistry, materials science, and synthetic biology, a wider range of material, chemical and biotic building blocks are emerging, such as metamaterials and active matter (Silva et al., 2014; Bernheim-Groswasser et al., 2018; McGivern, 2019; De Nicola et al., 2020; Pishvar and Harne, 2020; Zhang et al., 2020), novel chemical compounds (Gromski et al., 2020), and computer-designed organisms (Kriegman et al., 2020). These new building blocks may in turn allow artificial or natural evolutionary pressures to design hybrid systems that set new performance records for speed, dexterity, metabolic efficiency, or intelligence, while easing unsatisfying metabolic, biomechanical and adaptive tradeoffs. Machine interfaces are also being used to connect brains into novel compound entities, enhancing performance and collaboration (Jiang et al., 2019). If the net is cast wider, and virtual reality, the Internet of Things, and human societies are combined such that they create and co-create one another, it may be possible to obtain the best of both, of all worlds. This would not be the purely mechanistic World-Machine that Newton originally envisioned, but closer to the transhumanist ideal of a more perfect union of technology, biology, and society.

CONCLUSION

Living cells and tissues are not really machines; but then again, nothing is *really* anything – all metaphors are wrong, but some are more useful than others. If we update the machine metaphor in biology in accordance with modern research in the science of machine behavior, it can help deepen conceptual understanding and drive empirical research in ways that siloed efforts based on prior centuries' facile distinctions cannot. If we do not take this journey, we will not only be left mute in the face of numerous hybrid creatures in which these two supposedly different world interact tightly but will also have greatly limited our ability to design and control complex systems that could address many needs of individuals and society as a whole.

Are living things a computer (Wang and Gribskov, 2005; Bray, 2009)? It is a popular trope that humans naively seek to understand mind and life via the common engineering metaphors of the age - hydraulics, gears, electric circuits. However, this easy criticism, suggesting myopia and hyperfocus on each era's shiny new technology, is mistaken. The reason such technologies are compelling is that they are showing us the space of what is possible, by exploring newly discovered laws of nature in novel configurations. Are cells like steam engines? Not overtly, but the laws of thermodynamics that steam engines helped us to uncover and exploit are as important for biology as they are for physics. Cells and tissues are certainly not like the computers many of us use today, but that critique misses the point. Today's familiar computers are but a tiny portion of the huge space of systems that compute, and in this deeper, more important sense, living things are profitably studied with the deep concepts of computer science. Computer science offers many tools to help make more profound our understanding of the relationship between "minds and bodies" – physical structures that facilitate and constrain robustness, plasticity, memory, planning, intelligence, and all of the other key features of life.

It is now essential to re-draw (or perhaps erase) artificial boundaries between biology and engineering; the tight separation of disciplines is a hold-over from a past age, and is not the right way to carve nature by its joints. We live in a universe containing a rich, continuous option space of agents with which we can interact by re-wiring, training, motivating, signaling, communicating, and persuading. A better synergy between life sciences and engineering helps us to understand graded agency and nano-cognition across levels in biology, and create new instances (Pattee, 1979, 1982, 1989, 2001; Baluška and Levin, 2016). Indeed, biology and computer science are not two different fields; they are both branches of information science, working in distinct media with much in common. The science of behavior, applied to embodied computation in physical media that can be evolved or designed or both, is a new emerging field that will help us map and explore the enormous and fascinating space of possible machines across many scales of autonomy and composition. At stake is a most exciting future: where deep understanding of the origins and possible embodiments of autonomy help natural and synthetic systems reach their full potential.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

Both authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

We thank the members of our labs and helpful members of the Twitterverse for many useful discussions and pointers to literature. We also thank Santosh Manicka and Vaibhav Pai for comments on early versions of this manuscript. ML gratefully acknowledges support of the Barton Family Foundation, the Elisabeth Giauque Trust, and the Templeton World Charity Foundation. JB wishes to thank the family of Cyril G. Veinott for their financial support, as well as the University of Vermont's Office of the Vice President for Research.

REFERENCES

Alcala-Zermeno, J. J., Gregg, N. M., Van Gompel, J. J., Stead, M., Worrell, G. A., and Lundstrom, B. N. (2020). Cortical and thalamic electrode implant followed by temporary continuous subthreshold stimulation yields long-term seizure freedom: a case report. *Epilepsy Behav. Rep.* 14:100390. doi: 10.1016/j.ebr.2020.100390

Ando, N., and Kanzaki, R. (2020). Insect-machine hybrid robot. *Curr. Opin Insect. Sci.* 42, 61–69. doi: 10.1016/j.cois.2020.09.006

Applewhite, P. B. (1975). "Plant and animal behavior: an introductory comparison," in *Aneural Organisms in Neurobiology*, ed. E. M. Eisenstein (New York: Plenum Press), 131–139. doi: 10.1007/978-1-4613-4473-5_9

Arbib, M. (1961). *Turing Machines, Finite Automata and Neural Nets*. New York, NY: ACM.

Bacchus, W., Aubel, D., and Fussenegger, M. (2013). Biomedically relevant circuit-design strategies in mammalian synthetic biology. *Mol. Syst. Biol.* 9:691. doi: 10.1038/msb.2013.48

Bach-y-Rita, P. (1967). Sensory plasticity. applications to a vision substitution system. *Acta Neurol. Scand.* 43, 417–426.

Baluška, F., and Levin, M. (2016). On having no head: cognition throughout biological systems. *Front. Psychol.* 7:902.

Barbieri, M. (1998). The organic codes. the basic mechanism of macroevolution. *Riv. Biol.* 91, 481–513.

Barbieri, M. (2018). What is code biology? *Biosystems* 164, 1–10. doi: 10.1016/j.biosystems.2017.10.005

Barbieri, M. (2019). A general model on the origin of biological codes. *Biosystems* 181, 11–19. doi: 10.1016/j.biosystems.2019.04.010

Beer, R. D. (2004). Autopoiesis and cognition in the game of life. *Artif. Life* 10, 309–326. doi: 10.1162/1064546041255539

Beer, R. D. (2014). The cognitive domain of a glider in the game of life. *Artif. Life* 20, 183–206. doi: 10.1162/artl_a_00125

Beer, R. D. (2015). Characterizing autopoiesis in the game of life. *Artif. Life* 21, 1–19. doi: 10.1162/artl_a_00143

Beer, R. D., and Williams, P. L. (2015). Information processing and dynamics in minimally cognitive agents. *Cogn. Sci.* 39, 1–38. doi: 10.1111/cogs.12142

Beinhocker, E. D. (2020). *The Origin of Wealth: Evolution, Complexity, and the Radical Remaking of Economics*. Boston, MA: Harvard Business Press.

Belousov, L. V. (2008). "Our standpoint different from common." (scientific heritage of alexander gurwitsch) [English]. *Russ. J. Dev. Biol.* 39, 307–315. doi: 10.1134/s1062360408050081

Bernatskiy, A., and Bongard, J. (2017). Choice of robot morphology can prohibit modular control and disrupt evolution. in *Proceedings of the Fourteenth European Conference on Artificial Life*. Switzerland: ECAL.

Bernheim-Groswasser, A., Gov, N. S., Safran, S. A., and Tzlil, S. (2018). Living matter: mesoscopic active materials. *Adv. Mater.* 30:e1707028.

Bertschinger, N., Olbrich, E., Ay, N., and Jost, J. (2008). Autonomy: an information theoretic perspective. *Bio. Syst.* 91, 331–345. doi: 10.1016/j.biosystems.2007.05.018

Birnbaum, K. D., and Alvarado, A. S. (2008). Slicing across kingdoms: regeneration in plants and animals. *Cell* 132, 697–710. doi: 10.1016/j.cell.2008.01.040

Biswas, S., Manicka, S., Hoel, E., and Levin, M. (2021). Gene regulatory networks exhibit several kinds of memory: quantification of memory in biological and random transcriptional networks. *iScience* 102131. doi: 10.1016/j.isci.2021.102131

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, (Barcelona: Spain), 4349–4357.

Bongard, J. C. (2013). Evolutionary robotics. *Commun. ACM* 56, 74–83.

Bongard, J., Zykov, V., and Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science* 314, 1118–1121. doi: 10.1126/science.1133687

Bozorgzadeh, B., Schuweiler, D. R., Bobak, M. J., Garris, P. A., and Mohseni, P. (2016). Neurochemostat: a neural interface soc with integrated chemometrics for closed-loop regulation of brain dopamine. *IEEE Trans. Biomed. Circ. Syst.* 10, 654–667. doi: 10.1109/tbcas.2015.2453791

Braitenberg, V. (1984). *Vehicles, Experiments in Synthetic Psychology*. Cambridge: MIT Press.

Bray, D. (2009). *Wetware: a Computer in Every Living Cell*. New Haven: Yale University Press.

Brodbeck, L., Hauser, S., and Iida, F. (2018). "Robotic invention: challenges and perspectives for model-free design optimization of dynamic locomotion robots," in *Robotics Research*, eds A. Bicchi and W. Burgard (Cham: Springer), 581–596. doi: 10.1007/978-3-319-60916-4_33

Bronfman, Z. Z., Ginsburg, S., and Jablonka, E. (2016). The transition to minimal consciousness through the evolution of associative learning. *Front. Psychol.* 7:1954.

Calabretta, R., Nolfi, S., Parisi, D., and Wagner, G. P. (2000). Duplication of modules facilitates the evolution of functional specialization. *Artif. Life* 6, 69–84. doi: 10.1162/106454600568320

Calvo, P., Sahi, V. P., and Trewavas, A. (2017). Are plants sentient? *Plant Cell Environ.* 40, 2858–2869.

Chamola, V., Vineet, A., Nayyar, A., and Hossain, E. (2020). Brain-computer interface-based humanoid control: a review. *Sensors* 20:3620. doi: 10.3390/s20133620

Chatterjee, G., Dalchau, N., Muscat, R. A., Phillips, A., and Seelig, G. (2017). A spatially localized architecture for fast and modular DNA computing. *Nat. Nanotechnol.* 12, 920–927. doi: 10.1038/nnano.2017.127

Chen, J. P., Rogers, L. G., Anderson, L., Andrews, U., Brzoska, A., and Coffey, A. (2017). Power dissipation in fractal AC circuits. *J. Phys. Math. Theor.* 50:325205. doi: 10.1088/1751-8121/aa7a66

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis* 58, 7–19.

Clune, J., Mouret, J. B., and Lipson, H. (2013). The evolutionary origins of modularity. *Proc. Biol. Sci.* 280:20122863. doi: 10.1098/rspb.2012.2863

Conrad, M. (1989). The brain-machine disanalogy. *Biosystems* 22, 197–213. doi: 10.1016/0303-2647(89)90061-0

Corucci, F., Cheney, N., Lipson, H., Laschi, C., and Bongard, J. C. (2015). "Material properties affect evolution's ability to exploit morphological computation in growing soft-bodied creatures," in *Proceedings of The Fifteenth International Conference on the Synthesis and Simulation of Living Systems*, (Cambridge: MIT press).

Cruse, H., and Schilling, M. (2013). How and to what end may consciousness contribute to action? attributing properties of consciousness to an embodied, minimally cognitive artificial neural network. *Front. Psychol.* 4:324.

Cvrckova, F., Zarsky, V., and Markos, A. (2016). Plant studies may lead us to rethink the concept of behavior. *Front. Psychol.* 7:622.

Danilov, Y., and Tyler, M. (2005). Brainport: an alternative input to the brain. *J. Integr. Neurosci.* 4, 537–550.

Davidson, L. A. (2012). Epithelial machines that shape the embryo. *Trends Cell Biol.* 22, 82–87. doi: 10.1016/j.tcb.2011.10.005

De Nicola, F., Puthiya Purayil, N. S., Miseikis, V., Spirito, D., Tomadin, A., and Coletti, C. (2020). Graphene plasmonic fractal metamaterials for broadband photodetectors. *Sci. Rep.* 10:6882.

DeMarse, T. B., and Dockendorf, K. P. (2005). *Adaptive Flight Control with Living Neuronal Networks on Microelectrode Arrays*. Piscataway, NJ: IEEE.

Dennett, D. (1987). *The Intentional Stance*. Cambridge, Mass: MIT Press.

Dennett, D. C. (2017). *From Bacteria to Bach and Back : the Evolution of Minds*. New York: W.W. Norton & Company.

Diaspro, A. (2004). Introduction: a nanoworld under the microscope–from cell trafficking to molecular machines. *Microsc. Res. Tech.* 65, 167–168. doi: 10.1002/jemt.20137

Durant, F., Morokuma, J., Fields, C., Williams, K., Adams, D. S., and Levin, M. (2017). Long-term, stochastic editing of regenerative anatomy via targeting endogenous bioelectric gradients. *Biophys. J.* 112, 2231–2243. doi: 10.1016/j.bpj.2017.04.011

Fields, C., and Levin, M. (2017). Multiscale memory and bioelectric error correction in the cytoplasm–cytoskeleton-membrane system. *Wiley Interdiscipl. Rev. Syst. Biol. Med.* 10:e1410–n/a. doi: 10.1002/wsbm.1410

Fields, C., and Levin, M. (2018). Are planaria individuals? what regenerative biology is telling us about the nature of multicellularity. *Evol. Biol.* 45, 237–247. doi: 10.1007/s11692-018-9448-9

Fields, C., and Levin, M. (2020). Scale-free biology: integrating evolutionary and developmental thinking. *BioEssays* 42:e1900228. doi: 10.1002/bies.201900228

Froese, T., McGann, M., Bigge, W., Spiers, A., and Seth, A. K. (2012). The enactive torch: a new tool for the science of perception. *Ieee T Haptics* 5, 365–375. doi: 10.1109/TOH.2011.57

Fuchslin, R. M., Dzyakanchuk, A., Flumini, D., Hauser, H., Hunt, K. J., and Luchsinger, R. (2013). Morphological computation and morphological control: steps toward a formal theory and applications. *Artif. Life* 19, 9–34.

Garcia, S., and Trinh, C. T. (2019). Modular design: implementing proven engineering principles in biotechnology. *Biotechnol. Adv.* 37:107403. doi: 10.1016/j.biotechadv.2019.06.002

Garzon, P. C., and Keijzer, F. (2011). Plants: adaptive behavior, root-brains, and minimal cognition. *Adapt. Behav.* 19, 155–171. doi: 10.1177/1059712311409446

Gawne, R., McKenna, K. Z., and Levin, M. (2020). Competitive and coordinative interactions between body parts produce adaptive developmental outcomes. *BioEssays* 42:e1900245. doi: 10.1002/bies.201900245

Gilbert, S. F., and Sarkar, S. (2000). Embracing complexity: organicism for the 21st century. *Dev. Dyn.* 219, 1–9. doi: 10.1002/1097-0177(2000)9999:9999<::AID-DVDY1036>3.0.CO;2-A

Goodwin, B. C. (1977). Cognitive biology. *Commun. Cogn.* 10, 87–91.

Goodwin, B. C. (1978). A cognitive view of biological process. *J. Soc. Biol. Struct.* 1, 117–125. doi: 10.1016/S0140-1750(78)80001-3

Goodwin, B. C. (2000). The life of form. emergent patterns of morphological transformation. comptes rendus de l'Academie des sciences. *Serie III Sci. de la vie* 323, 15–21. doi: 10.1016/S0764-4469(00)00107-4

Green, A. M., and Kalaska, J. F. (2011). Learning to move machines with the mind. *Trends Neurosci.* 34, 61–75.

Gromski, P. S., Granda, J. M., and Cronin, L. (2020). Universal chemical synthesis and discovery with 'the chemputer'. *Trends Chem.* 2, 4–12. doi: 10.1016/j.trechm.2019.07.004

Grosenick, L., Marshel, J. H., and Deisseroth, K. (2015). Closed-loop and activity-guided optogenetic control. *Neuron* 86, 106–139. doi: 10.1016/j.neuron.2015.03.034

Guidotti, R., Monreale, A., Ruggieri, S., Turin, F., Giannotti, F., and Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Comput. Surv.* 51:93.

Gurwitsch, A. G. (1944). *A Biological Field Theory*. Moscow: Nauka.

Haensch, W., Gokmen, T., and Puri, R. (2018). *The Next Generation of Deep Learning Hardware: Analog Computing*. Piscataway, NJ: IEEE.

Hamann, H., Wahby, M., Schmickl, T., Zahadat, P., Hofstadler, D., and Stoy, K. (2015). "flora robotica - mixed societies of symbiotic robot-plant bio-hybrids," in *Proceedings of the 2015 IEEE Symposium Series on Computational Intelligence*, (Piscataway: IEEE), 1102–1109.

Hawkes, E., An, B., Benbernou, N. M., Tanaka, H., Kim, S., and Demaine, E. D. (2010). Programmable matter by folding. *Proc. Natl. Acad. Sci. U.S.A.* 107, 12441–12445.

Hiett, P. J. (1999). Characterizing critical rules at the 'edge of chaos'. *Bio. Syst.* 49, 127–142. doi: 10.1016/S0303-2647(98)00039-2

Ho, W. M., and Fox, S. W. (1988). *Evolutionary Processes and Metaphors*. New York: Wiley.

Hoffmeyer, J. (2000). Code-duality and the epistemic cut. *Ann. N. Y. Acad. Sci.* 901, 175–186. doi: 10.1111/j.1749-6632.2000.tb06277.x

Honeck, R. P., and Hoffman, R. R. (1980). *Cognition and Figurative Language*. Hillsdale, N.J: L. Erlbaum Associates.

Jiang, L., Stocco, A., Losey, D. M., Abernethy, J. A., Prat, C. S., and Rao, P. N. (2019). BrainNet: a multi-person brain-to-brain interface for direct collaboration between brains. *Sci. Rep.* 9:6115. doi: 10.1038/s41598-019-41895-7

Jonas, E., and Kording, K. (2016). Could a neuroscientist understand a microprocessor? *biooRxiv [preprint]* doi: 10.1371/journal.pcbi.1005268

Jones, R., Haufe, P., Sells, E., Iravani, P., Olliver, V., Palmer, C., et al. (2011). RepRap - the replicating rapid prototyper. *Robotica* 29, 177–191. doi: 10.1017/S026357471000069X

Kamm, R. D., and Bashir, R. (2014). Creating living cellular machines. *Ann. Biomed. Eng.* 42, 445–459. doi: 10.1007/s10439-013-0902-7

Kamm, R. D., Bashir, R., Arora, N., Dar, R. D., Gillette, M. U., and Griffith, L. G. (2018). Perspective: the promise of multi-cellular engineered living systems. *APL Bioeng.* 2:040901.

Kauffman, S. A., and Johnsen, S. (1991). Coevolution to the edge of chaos: coupled fitness landscapes, poised states, and coevolutionary avalanches. *J. Theor. Biol.* 149, 467–505. doi: 10.1016/S0022-5193(05)80094-3

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optiization. *arXiv [preprint]*.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005). *Empowerment: a Universal Agent-Centric Measure of Control*. Piscataway, NJ: IEEE.

Kriegman, S., Blackiston, D., Levin, M., and Bongard, J. (2020). A scalable pipeline for designing reconfigurable organisms. *Proc. Natl. Acad. Sci. U.S.A.* 117, 1853–1859. doi: 10.1073/pnas.1910837117

Kriegman, S., Walker, S., Shah, D. S., Levin, M., Kramer-Bottiglio, R., and Bongard, J. (2019). *Automated Shapeshifting for Function Recovery in Damaged Robots*. Freiburg im Breisgau: Robotics: Science and Systems XV, 28.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Kwiatkowski, R., and Lipson, H. (2019). Task-agnostic self-modeling machines. *Sci. Robot.* 4:eaau9354. doi: 10.1126/scirobotics.aau9354

Langton, C. G. (1995). *Artificial life : an Overview*. Cambridge, Mass: MIT Press. doi: 10.7551/mitpress/1427.001.0001

Lee, S. H., Piao, H., Cho, Y. C., Kim, S. N., Choi, G., and Kim, C. R. (2019). Implantable multireservoir device with stimulus-responsive membrane for on-demand and pulsatile delivery of growth hormone. *Proc. Natl. Acad. Sci. U.S.A.* 116, 11664–11672.

Lehman, J., and Stanley, K. O. (2011). "Novelty search and the problem with objectives," in *Genetic Programming Theory and Practice*, eds R. Riola, E. Vladislavleva, and J. H. Moore (Berlin: Springer), 37–56.

Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., and Beaulieu, J. (2020). The surprising creativity of digital evolution: a collection of anecdotes from the evolutionary computation and artificial life research communities. *Artif. Life* 26, 274–306.

Levin, M. (2019). The computational boundary of a "self": developmental bioelectricity drives multicellularity and scale-free cognition. *Front. Psychol.* 10:2688. doi: 10.3389/fpsyg.2019.02688

Levin, M. (2020a). Life, death, and self: fundamental questions of primitive cognition viewed through the lens of body plasticity and synthetic organisms. *Biochem. Biophys. Res. Commun.* doi: 10.1016/j.bbrc.2020.10.077 Online ahead of print.

Levin, M. (2020b). The biophysics of regenerative repair suggests new perspectives on biological causation. *BioEssays* 42:e1900146. doi: 10.1002/bies.201900146

Levin, M., and Martyniuk, C. J. (2018). The bioelectric code: an ancient computational medium for dynamic control of growth and form. *Biosystems* 164, 76–93. doi: 10.1016/j.biosystems.2017.08.009

Levin, M., Bongard, J., and Lunshof, J. E. (2020). Applications and ethics of computer-designed organisms. *Nat. Rev. Mol. Cell Biol.* 21, 655–656. doi: 10.1038/s41580-020-00284-z

Levin, M., Pietak, A. M., and Bischof, J. (2018). Planarian regeneration as a model of anatomical homeostasis: recent progress in biophysical and computational approaches. *Semin. Cell Dev. Biol.* 87, 125–144. doi: 10.1016/j.semcdb.2018.04.003

Lewis, A. C. F. (2020). Where bioethics meets machine ethics. *Am. J. Bioeth.* 20, 22–24. doi: 10.1080/15265161.2020.1819471

Liu, X., Ramirez, S., and Tonegawa, S. (2014). Inception of a false memory by optogenetic manipulation of a hippocampal memory engram. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369:20130142. doi: 10.1098/rstb.2013.0142

Lobo, D., Solano, M., Bubenik, G. A., and Levin, M. (2014). A linear-encoding model explains the variability of the target morphology in regeneration. *J. R. Soc.* 11:20130918.

Lucas, J. R. (1961). Minds, machines, and godel. *Philosophy* 36, 112–127. doi: 10.1017/S0031819100057983

Lyon, P. (2006). The biogenic approach to cognition. *Cogn. Process.* 7, 11–29. doi: 10.1007/s10339-005-0016-8

Ma, Z., Stork, T., Bergles, D. E., and Freeman, M. R. (2016). Neuromodulators signal through astrocytes to alter neural circuit activity and behaviour. *Nature* 539, 428–432.

Man, K., and Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nat. Mach. Intell.* 1, 446–452. doi: 10.1038/s42256-019-0103-7

Manicka, S., and Levin, M. (2019). The cognitive lens: a primer on conceptual tools for analysing information processing in developmental and regenerative morphogenesis. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 374:20180369. doi: 10.1098/rstb.2018.0369

Mariscal, C., and Doolittle, W. F. (2020). Life and life only: a radical alternative to life definitionism. *Synthese* 197, 2975–2989. doi: 10.1007/s11229-018-1852-2

Marr, D. (1982). *Vision : a Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: Freeman, W. H.

McGivern, P. (2019). Active materials: minimal models of cognition? *Adapt. Behav.* 28:105971231989174. doi: 10.1177/1059712319891742

McLennan-Smith, T. A., Roberts, D. O., and Sidhu, H. S. (2020). Emergent behavior in an adversarial synchronization and swarming model. *Phys Rev. E* 102:032607. doi: 10.1103/PhysRevE.102.032607

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., and Bellemare, M. G. (2015). Human-level control through deep reinforcement learning. *Nature* 518, 529–533.

Montevil, M., and Mossio, M. (2015). Biological organisation as closure of constraints. *J. Theor. Biol.* 372, 179–191. doi: 10.1016/j.jtbi.2015.02.029

Moore, C. I., and Cao, R. (2008). The hemo-neural hypothesis: on the role of blood flow in information processing. *J. Neurophysiol.* 99, 2035–2047. doi: 10.1152/jn.01366.2006

Mora, T., and Bialek, W. (2011). Are biological systems poised at criticality? *J. Stat. Phys.* 144, 268–302. doi: 10.1007/s10955-011-0229-4

Nakajima, K., Hauser, H., Li, T., and Pfeifer, R. (2015). Information processing via physical soft body. *Sci. Rep.* 5:10487.

Newman, J. P., Fong, M. F., Millard, D. C., Whitmire, C. J., Stanley, G. B., and Potter, S. M. (2015). Optogenetic feedback control of neural activity. *Elife* 4:e07192.

Nicholson, D. J. (2012). The concept of mechanism in biology. *Stud. Hist. Philos. Biol. Biomed. Sci.* 43, 152–163.

Nicholson, D. J. (2013). Organisms not equal machines. *Stud. Hist. Philos. Biol. Biomed. Sci.* 44, 669–678.

Nicholson, D. J. (2014). The machine conception of the organism in development and evolution: a critical analysis. *Stud. Hist. Philos. Biol. Biomed. Sci.* 48, 162–174. doi: 10.1016/j.shpsc.2014.08.003

Nicholson, D. J. (2019). Is the cell really a machine? *J. Theor. Biol.* 477, 108–126. doi: 10.1016/j.jtbi.2019.06.002

Nitsch, V., and Popp, M. (2014). Emotions in robot psychology. *Biol. Cybern.* 108, 621–629. doi: 10.1007/s00422-014-0594-6

Noble, D. (2012). A theory of biological relativity: no privileged level of causation. *Interface Focus* 2, 55–64. doi: 10.1098/rsfs.2011.0067

Oudeyer, P. Y., and Kaplan, F. (2007). What is intrinsic motivation? a typology of computational approaches. *Front. Neurorobot.* 1:6. doi: 10.3389/neuro.12.006.2007

Pais-Vieira, M., Chiuffa, G., Lebedev, M., Yadav, A., and Nicolelis, M. A. (2015). Building an organic computing device with multiple interconnected brains. *Sci. Rep.* 5:11869. doi: 10.1038/srep14937

Pashaie, R., Baumgartner, R., Richner, T. J., Brodnick, S. K., Azimipour, M., Eliceiri, K. W., et al. (2015). Closed-loop optogenetic brain interface. *IEEE Trans. Biomed. Eng.* 62, 2327–2337.

Pattee, H. H. (1979). The complementarity principle and the origin of macromolecular information. *Biosystems* 11, 217–226. doi: 10.1016/0303-2647(79)90013-3

Pattee, H. H. (1982). Cell psychology: an evolutionary approach to the symbol-matter problem. *Cogn. Brain Theory* 5, 325–341.

Pattee, H. H. (1989). The measurement problem in artificial world models. *Biosystems* 23, 281–289. doi: 10.1016/0303-2647(89)90036-1

Pattee, H. H. (2001). The physics of symbols: bridging the epistemic cut. *Biosystems* 60, 5–21. doi: 10.1016/S0303-2647(01)00104-6

Pavlic, T. P., and Pratt, S. C. (2013). "Superorganismic behavior via human computation," in *Handbook of Human COmputtion*, ed. P. Michelucci (Arizona State University: Tempe). doi: 10.1007/978-1-4614-8806-4_74

Pezzulo, G., and Levin, M. (2015). Re-membering the body: applications of computational neuroscience to the top-down control of regeneration of limbs and other complex organs. *Integr. Biol.* 7, 1487–1517. doi: 10.1039/C5IB00221D

Pezzulo, G., and Levin, M. (2016). Top-down models in biology: explanation and control of complex living systems above the molecular level. *J. R. Soc. Interface* 13:20160555. doi: 10.1098/rsif.2016.0555

Pfeifer, R., and Bongard, J. (2006). *How the Body Shapes the Way We Think: a New View of Intelligence*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/3585.001.0001

Pishvar, M., and Harne, R. L. (2020). Foundations for soft, smart matter by active mechanical metamaterials. *Adv. Sci.* 7:2001384. doi: 10.1002/advs.202001384

Powers, J., Grindle, R., Kriegman, S., Frati, L., Cheney, N., et al. (2020). "Morphology dictates learnability in neural controllers," in *Proceedings of the ALIFE 2020: The 2020 Conference on Artificial Life*, (Cambridge: MIT Press). doi: 10.1162/isal_a_00243

Preskill, J. (2018). Quantum computing in the NISQ era and beyond. *Quantum* 2:79. doi: 10.22331/q-2018-08-06-79

Ptito, M., Moesgaard, S. M., Gjedde, A., and Kupers, R. (2005). Cross-modal plasticity revealed by electrotactile stimulation of the tongue in the congenitally blind. *Brain J. Neurol.* 128, 606–614. doi: 10.1093/brain/awh380

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., and Breazeal, C. (2019). Machine behaviour. *Nature* 568, 477–486.

Ramirez, S., Liu, X., Lin, P. A., Suh, J., Pignatelli, M., and Redondo, R. L. (2013). Creating a false memory in the hippocampus. *Science* 341, 387–391.

Reger, B. D., Fleming, K. M., Sanguineti, V., Alford, S., and Mussa-Ivaldi, F. A. (2000). Connecting brains to robots: an artificial body for studying the computational properties of neural tissues. *Artif. Life* 6, 307–324. doi: 10.1162/106454600300103656

Rosen, R. (1974). Biological-systems as organizational paradigms. *Int. J. Gen. Syst.* 1, 165–174. doi: 10.1080/03081077408960769

Rosen, R. (1985). *Anticipatory Systems : Philosophical, Mathematical, and Methodological Foundations*. New York: Pergamon Press.

Roy, D. S., Kitamura, T., Okuyama, T., Ogawa, S. K., Sun, C., Obata, Y., et al. (2017). Distinct neural circuits for the formation and retrieval of episodic memories. *Cell* 170, 1000-1012 e19.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi: 10.1038/s42256-019-0048-x

Salvi, C., Beeman, M., Bikson, M., McKinley, R., and Grafman, J. (2020). TDCS to the right anterior temporal lobe facilitates insight problem-solving. *Sci. Rep.* 10:946. doi: 10.1038/s41598-020-57724-1

Sampaio, E., Maris, S., and Bach-y-Rita, P. (2001). Brain plasticity: 'visual' acuity of blind persons via the tongue. *Brain Res.* 908, 204–207. doi: 10.1016/S0006-8993(01)02667-1

Santello, M., Toni, N., and Volterra, A. (2019). Astrocyte function from information processing to cognition and cognitive impairment. *Nat. Neurosci.* 22, 154–166. doi: 10.1038/s41593-018-0325-8

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., and Jain, B. (2019). "Habitat: a platform for embodied ai research," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (Piscataway, NJ: IEEE), 9339–9347.

Schmidt, M., and Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science* 324, 81–85. doi: 10.1126/science.1165893

Schulkin, J., and Sterling, P. (2019). Allostasis: a brain-centered, predictive mode of physiological regulation. *Trends Neurosci.* 42, 740–752. doi: 10.1016/j.tins.2019.07.010

Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proc. Natl. Acad. Sci. U.S.A.* 117, 30033–30038. doi: 10.1073/pnas.1907373117

Semenov, A. S. (2020). *Essentials of Fractal Programming, Advances in Theory and Practice of Computational Mechanics*. Singapore: Springer, 373–386. doi: 10.1007/978-981-15-2600-8_25

Shah, D. S., Powers, J. P., Tilton, L. G., Kriegman, S., Bongard, J., and Kramer-Bottiglio, R. (2020). A soft robot that adapts to environments through shape change. *Nat. Mach. Intell.* 3, 51–59. doi: 10.1038/s42256-020-00263-1

Shanechi, M. M., Hu, R. C., and Williams, Z. M. (2014). A cortical-spinal prosthesis for targeted limb movement in paralysed primate avatars. *Nat. Commun.* 5:3237. doi: 10.1038/ncomms4237

Shen, G., Dwivedi, K., Majima, K., Horikawa, T., and Kamitani, Y. (2019a). End-to-end deep image reconstruction from human brain activity. *Front. Comput. Neurosci.* 13:21.

Shen, G., Horikawa, T., Majima, K., and Kamitani, Y. (2019b). Deep image reconstruction from human brain activity. *PLoS Comput. Biol.* 15:e1006633. doi: 10.1371/journal.pcbi.1006633

Silva, A., Monticone, F., Castaldi, G., Galdi, V., Alu, A., and Engheta, N. (2014). Performing mathematical operations with metamaterials. *Science* 343, 160–163.

Slavkov, I., Carrillo-Zapata, D., Carranza, N., Diego, X., Jansson, F., Kaandorp, J., et al. (2018). Morphogenesis in robot swarms. *Sci. Robot.* 3:eaau9178. doi: 10.1126/scirobotics.aau9178

Slusarczyk, A. L., Lin, A., and Weiss, R. (2012). Foundations for the design and implementation of synthetic genetic circuits. *Nat. Rev. Genet.* 13, 406–420. doi: 10.1038/nrg3227

Solé, R. V., and Goodwin, B. C. (2000). *Signs of Life : How Complexity Pervades Biology*. New York: Basic Books.

Sole, R., Amor, D. R., Duran-Nebreda, S., Conde-Pueyo, N., Carbonell-Ballestero, M., and Montanez, R. (2016). Synthetic collective intelligence. *Biosystems* 148, 47–61. doi: 10.1016/j.biosystems.2016.01.002

Sporns, O., Tononi, G., and Edelman, G. M. (2000). Theoretical neuroanatomy: relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cereb Cortex* 10, 127–141. doi: 10.1093/cercor/10.2.127

Suthana, N., Haneef, Z., Stern, J., Mukamel, R., Behnke, E., Knowlton, B., et al. (2012). Memory enhancement and deep-brain stimulation of the entorhinal area. *N. Engl. J. Med.* 366, 502–510.

Swain, D. T., Couzin, I. D., and Leonard, N. E. (2012). Real-time feedback-controlled robotic fish for behavioral experiments with fish schools. *Proc. IEEE* 100, 150–163. doi: 10.1109/JPROC.2011.2165449

Thomson, D., Zilkie, A., Bowers, J. E., Komljenovic, T., Reed, G. T., and Vivien, L. (2016). Roadmap on silicon photonics. *J. Optics* 18:073003.

Thubagere, A. J., Li, W., Johnson, R. F., Chen, Z., Doroudi, S., and Lee, Y. L. (2017). A cargo-sorting DNA robot. *Science* 357:eaan6558.

Trewavas, A. (2009). What is plant behaviour? *Plant Cell Environ.* 32, 606–616. doi: 10.1111/j.1365-3040.2009.01929.x

Turing, A. M. (1950). Computing machinery and intelligence. *Mind* 59, 433–460. doi: 10.1093/mind/LIX.236.433

Valentini, G., Moore, D. G., Hanson, J. R., Pavlic, T. P., Pratt, S. C., et al. (2018). "Transfer of information in collective decisions by artificial agents," in *Proceedings of the 2018 Conference on Artificial Life*, (Cambridge, MA: MIT press), 641–648.

Vandenberg, L. N., Adams, D. S., and Levin, M. (2012). Normalized shape and location of perturbed craniofacial structures in the Xenopus tadpole reveal an innate ability to achieve correct morphology. *Dev. Dyn.* 241, 863–878. doi: 10.1002/dvdy.23770

Varela, F. G., Maturana, H. R., and Uribe, R. (1974). Autopoiesis: the organization of living systems, its characterization and a model. *Biosystem* 5, 187–196.

Varela, F., and Maturana, H. (1972). Mechanism and biological explanation. *Philos. Sci.* 39, 378–382. doi: 10.1086/288458

Vernon, D., Lowe, R., Thill, S., and Ziemke, T. (2015). Embodied cognition and circular causality: on the role of constitutive autonomy in the reciprocal coupling of perception and action. *Front. Psychol.* 6:1660. doi: 10.3389/fpsyg.2015.01660

Vetere, G., Tran, L. M., Moberg, S., Steadman, P. E., Restivo, L., and Morrison, F. G. (2019). Memory formation in the absence of experience. *Nat. Neurosci.* 22, 933–940.

von Mammen, S., Hamann, H., and Heider, M. (2016). "Robot gardens: an augmented reality prototype for plant-robot biohybrid systems," in *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*, ed. E. Kruijff (New York, NY: Association for Computing Machinery), 139–142. doi: 10.1145/2993369.2993400

von Neumann, V., and Burks, A. W. (1966). *Theory of Self-Reproducing Automata*. Champaign, IL: University of Illinois Press.

Walker, S. I., and Davies, P. C. (2013). The algorithmic origins of life. *J. R. Soc. Interface* 10:20120869.

Walter, G. (1950). An imitation of life. *Sci. Am.* 182. doi: 10.1038/scientificamerican0550-42

Wang, D., and Gribskov, M. (2005). Examining the architecture of cellular computing through a comparative study with a computer. *J. R. Soc. Interface* 2, 187–195. doi: 10.1098/rsif.2005.0038

Wilson, N. R., Schummers, J., Runyan, C. A., Yan, S. X., Chen, R. E., Deng, Y., et al. (2013). Two-way communication with neural networks in vivo using focused light. *Nat. Protoc.* 8, 1184–1203.

Yu, X., Wang, C., and Yu, D. (2019). Configuration optimization of the tandem cooling-compression system for a novel precooled hypersonic airbreathing engine. *Energy Convers. Manag.* 197:111827. doi: 10.1016/j.enconman.2019.111827

Zhang, T., Liu, Q., Dan, Y., Yu, S., Han, X., Dai, J., et al. (2020). Machine learning and evolutionary algorithm studies of graphene metamaterials for optimized plasmon-induced transparency. *Opt Express* 28, 18899–18916.

Zhou, A., Ouyang, J., Su, J., Zhang, S., and Yan, S. (2020). Multimodal optimisation design of product forms based on aesthetic evaluation. *Int. J. Arts Technol.* 12, 128–154.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Computer Science

MINI REVIEW
published: 08 February 2022
doi: 10.3389/fcomp.2022.810358

The Brain-Computer Metaphor Debate Is Useless: A Matter of Semantics

Blake A. Richards[1,2,3,4]* *and Timothy P. Lillicrap*[5]

[1] Mila, Montreal, QC, Canada, [2] Department of Neurology and Neurosurgery, Montreal Neurological Institute, McGill University, Montreal, QC, Canada, [3] School of Computer Science, McGill University, Montreal, QC, Canada, [4] Learning in Machines and Brains Program, CIFAR, Toronto, ON, Canada, [5] DeepMind Inc., London, United Kingdom

OPEN ACCESS

Edited by:
Giorgio Matassi,
FRE3498 Ecologie et dynamique des systèmes anthropisés (EDYSAN), France

Reviewed by:
Marcos Cramer,
Technical University Dresden, Germany
Michael Levin,
Tufts University, United States

****Correspondence:***
Blake A. Richards
blake.richards@mila.quebec

Specialty section:
This article was submitted to Theoretical Computer Science, a section of the journal Frontiers in Computer Science

Received: *06 November 2021*
Accepted: *17 January 2022*
Published: *08 February 2022*

Citation:
Richards BA and Lillicrap TP (2022) The Brain-Computer Metaphor Debate Is Useless: A Matter of Semantics.
Front. Comput. Sci. 4:810358.
doi: 10.3389/fcomp.2022.810358

It is commonly assumed that usage of the word “computer” in the brain sciences reflects a metaphor. However, there is no single definition of the word “computer” in use. In fact, based on the usage of the word “computer” in computer science, a computer is merely some physical machinery that can in theory compute any computable function. According to this definition the brain is literally a computer; there is no metaphor. But, this deviates from how the word “computer” is used in other academic disciplines. According to the definition used outside of computer science, “computers” are human-made devices that engage in sequential processing of inputs to produce outputs. According to this definition, brains are not computers, and arguably, computers serve as a weak metaphor for brains. Thus, we argue that the recurring brain-computer metaphor debate is actually just a semantic disagreement, because brains are either literally computers or clearly not very much like computers at all, depending on one’s definitions. We propose that the best path forward is simply to put the debate to rest, and instead, have researchers be clear about which definition they are using in their work. In some circumstances, one can use the definition from computer science and simply ask, what type of computer is the brain? In other circumstances, it is important to use the other definition, and to clarify the ways in which our brains are radically different from the laptops, smartphones, and servers that surround us in modern life.

Keywords: neuroscience, psychology, computer science, brains, computers, Turing machines, parallel distributed processing

1. INTRODUCTION

Computation has been a central feature of research in the brain sciences (neuroscience, psychology, and cognitive science) for decades. Papers in the brain sciences are full of references to algorithms, coding, and information processing (Diamant, 2008; Maass, 2016; Oteiza et al., 2017). At the same time, there is a long and continuing history of debate around these words (Maccormac, 1986; West and Travis, 1991; Smith, 1993; Vlasits, 2017). According to many scientists and philosophers, computers are used as a metaphor to understand brains and this metaphor can be misleading or counter-productive (Carello et al., 1984; Cisek, 1999; Epstein, 2016; Cobb, 2020). Throughout the history of the brain sciences over the last 80 years, one can find researchers who comfortably use computational theory and language to explore and understand brains (Marcus, 2015), as well as researchers who reject the use of such concepts for use with brains (Epstein, 2016). Indeed, the early dream of cognitive science in the second half of the twentieth century depended on the links between brain sciences and artificial intelligence (AI) (Newell, 1980; Simon, 1980; Pylyshyn, 1984; Hunt, 1989), yet the failure to make good progress in AI in the 1970's, 80's, and 90's, and the inability to connect such systems convincingly to the brain sciences, led some researchers to conclude that the "metaphor of the brain as a computer" was broken at its foundations (Dreyfus and Hubert, 1992; Van Gelder, 1998). To this day, one can still find in equal measure both brain scientists who use theories from computer science (Kwisthout and van Rooij, 2020) and brain scientists who argue against the brain as a computer metaphor (Brette, 2018).

However, closer inspection of the debates on this topic reveal a fundamental misunderstanding between the participants regarding the definition of the word "computer". Indeed, many of the entries in these debates do not grapple concretely with the definition of the word "computer" before declaring either way that the brain is or is not well-explained with computational theory. To actually resolve this debate, it is helpful to bring the definition of "computer" into clear focus.

Here, we argue that closer examination of the manner in which both computer scientists and non-computer scientists use the word "computer" indicates that there are at least two distinct definitions in operation: (1) A definition from computer science rooted in the formal concepts of computable functions and algorithms. (2) A definition from outside of computer science based on the electronic devices we use on a regular basis and how they operate. To make matters worse, some neuroscientists, cognitive scientists, and psychologists have a mixed familiarity with the formal concepts from computer science that underpin the first definition. This means that semantic debates stemming from misaligned definitions are particularly apt to emerge in the context of the brain sciences, leading to proponents on either side who seem irreconcilable.

In this article, we clarify these two distinct definitions. We show that if one adopts the definition from computer science, then the question is not whether computers are a good metaphor for brains, because brains arguably are *literally* computers based on this definition. In contrast, if one adopts the definition from outside of computer science then brains are not computers, and arguably, computers are a very poor metaphor for brains. Thus, the argument over whether or not computers are a good or bad metaphor for brains is actually just a matter of semantics. Under one definition, brains are literally computers, whereas under another, they are clearly not. There is, therefore, little utility in continuing these debates. We close on a prescription for the brain sciences. We suggest that the question for scientists should instead be: if we adopt the definition from computer science, then what kind of a computer are brains? For those using the definition from outside of computer science, they can be assured that their brains work in a very different way than their laptops and their smartphones—an important point to clarify as we seek to better understand how brains work.

2. MEANING AS USE

Before we discuss the different definitions of the word "computer", it is important that we clarify our approach to the definitions and meanings of words. In this paper, we adopt a perspective that focuses on the *use* of words for understanding their meaning, and thus, their definition. Therefore, we will avoid telling the reader that, for example, "computers are formally defined as *X*, and everyone must adopt this definition". Instead, we will draw the reader's attention to the ways in which the word "computer" *is in fact used* in contexts inside and outside of computer science, and proceed from there.

Briefly, the idea that we can best understand the meaning of a word by looking at its use in context has a long history in philosophy, perhaps best exemplified by the works of Ludwig Wittgenstein. Wittgenstein argued in the Philosophical Investigations that "*in most cases, the meaning of a word is its use*" (Wittgenstein, 1953). This idea flies in the face of many of our intuitive notions about how words work; like the young Wittgenstein, many of us tend to think about meaning in terms of correspondence, i.e., that "*individual words in language name objects and sentences are combinations of such names*" (Wittgenstein, 1953). But, in fact, meaning is crucially modified by context and use, rather than corresponding to particular objects, so the meaning of most words are fuzzy and impossible to write down precisely and uniquely. Wittgenstein showed how much confusion is generated by failing to pay attention to how words are used in context and we believe that much of the confusion around the question "Is the brain (like) a computer?" results from just this sort of confusion. In particular, the "computer as brain" debate often devolves into a semantic disagreement generated by a mismatch in expectations between two uses of the word "computer", which we will clarify below.

Of course, it should be noted that there can be many working definitions of the word "computer", but only two that are prominent and important in our context. The definition used by computer scientists is important because it underpins work in computational neuroscience and AI. And, at the same time, the definition used by academics outside of computer science is important because it's the one that most writers in the brain sciences intuitively reach for during these debates. As we'll see, someone operating with the computer science definition who says that the "brain is a computer" is certainly correct. Simultaneously, someone using the definition from outside of computer science who says that "the brain is not a computer and computers are not a good metaphor for brains" is also correct. Thus, unless the time is taken to clear up the question of usage, there's bound to be disagreement with little ground given by either side. As such, we must first explore these two distinct uses.

3. THE USE OF THE WORD "COMPUTER" INSIDE COMPUTER SCIENCE

Here we will provide an overview of the definition of the word "computer" based on the use of the word in computer science. As we will describe, this use-based definition partly relies on the formal definition of the word "algorithm". However, the definition of "computer" derived solely from the formal definition for "algorithm" is actually so broad as to be nearly meaningless. Nonetheless, the use of the word "computer" in computer science shows that computer scientists generally mean something more restrictive than the formal definitions would indicate. As we will show, the more restrictive, use-based definition is still applicable to brains.

3.1. The Formal Definition of "Algorithm" and The Church-Turing Thesis

Within computer science the formal definition for the word "algorithm" dates back to the early twentieth century, before the invention of modern computers and the discipline of computer science as it exists today. Back then, what would become computer science was essentially a branch of mathematics. Many mathematicians at the time were concerned with questions about a class of mathematical tools that they called "effective methods". An effective method is a finite recipe that one can follow mechanically to arrive at an answer to some mathematical problem (Copeland, 2020), e.g., long division is an effective method for solving division problems with arbitrarily large numbers. Today, we refer to effective methods as "algorithms". The intuitive definition of an algorithm is therefore as above: a finite recipe that one can follow mechanically to arrive at an answer to some problem (Cormen et al., 2009). But, we also have a formal definition thanks to the work of those early mathematicians. For example, in 1900, the mathematician David Hilbert put forward a set of 23 problems to be solved in the twentieth century, the 10th of which was "*Can we develop an algorithm for determining whether the roots of a polynomial function are integers?*" (Hilbert, 1902). Later, these types of questions were expanded in scope to larger questions such as the *Entscheidungsproblem*, which asks whether there is an algorithm for determining whether any given statement is valid within an axiomatic language (Hilbert and Ackermann, 1999).

It turned out that these mathematicians had stumbled onto a very deep set of problems. As they began to explore algorithms more and more, they started to wonder whether some problems in mathematics, including Hilbert's 10th problem and the *Entscheidungsproblem*, did not in fact have any solution. The way this is sometimes phrased is, are there problems that are not "decidable"? A problem is "decidable" if and only if there exists an algorithm for solving it (Cormen et al., 2009), and there was a growing realization that some problems were likely not decidable. Of course, mathematicians being mathematicians, they desired a proof that an algorithm didn't exist in such cases. The problem was that at the time the definition of an algorithm was the informal, intuitive definition above. Without a formal definition of the word "algorithm" it was impossible to prove that some problems were, in fact, not decidable.

This set the stage for the development of modern computer science as we know it today. A pair of mathematicians, Alonzo Church and Alan Turing, independently decided to try to develop a formal definition for "algorithm" for the sake of developing proofs related to the *Entscheidungsproblem* and decidability more broadly (Church, 1936a,b; Turing, 1936). Church invented a formal logical system he called lambda calculus, and defined an algorithm as anything that could be done with lambda calculus (Church, 1936a,b). Turing invented a mathematical construct known as a Turing machine, and defined an algorithm as anything that could be done with Turing machines (Turing, 1936). Both researchers used their definitions to show that there was no solution to the *Entscheidungsproblem*. As well, while the two researchers had developed what looked like very different definitions, they turned out to be mathematically equivalent (Turing, 1937). Continued work in computability theory, the branch of computer science and mathematics concerned with the study of decidable problems, has suggested that any attempt to formalize the intuitive definition of algorithm will end up being equivalent to lambda calculus and Turing machines (Cook, 1992, 2014; Copeland, 2020). As such, computer scientists today largely accept an idea known as **the Church-Turing thesis**, which states (very roughly), that any algorithm can be implemented *via* a Turing machine, i.e., it proposes that we accept Church and Turing's definitions as given (Copeland, 2020). Thus, when people seek a proof that there is no algorithm for some problem, they often do so by proving that you can't solve the problem with a Turing machine (Cook, 1992).

Importantly for the discussion here, the formal definition for an "algorithm" also gives rise to a formal conception of the word "computer". Specifically, computer scientists define a "computable function" to be any function whose values can be determined using an algorithm. A "computer" is then formally physical machinery that can implement algorithms in order to solve computable functions (though one may also take a slightly more expansive approach Copeland, 1997). It's worth noting that this conception of what a computer is makes no reference to human made artifacts, or electronics, or silicon chips, etc. And, if we think back to the use of the word "computer" at the point in history when Church and Turing were working, this makes a lot of sense: "Computers" at this time were people whose job was to sit down with pencil and paper and use effective methods (i.e., algorithms) to solve various problems (e.g., to integrate equations) (Grier, 2001). Clearly, these people were computers according to the definition above, because they were solving computable functions, even though they were of course not human-made artifacts. Thus, the formal definition of the word "algorithm" rests on the Church-Turing Thesis, and this in turn provides us with a formal definition of "computable functions", which is what "computers" solve. And, none of this has anything to do with the physical characteristics or internal workings of the computer, only with its ability to physically implement computable functions.

3.2. Limiting the Scope of the Formal Definition in Practice in Computer Science

If we consider the definition above for "computer", a problem arises: this definition can be applied to almost any object in the universe. Consider for a moment the fact that the movement of objects in the world can be described by computable functions, e.g., the parabolic curve of a thrown ball. As such, the definition that rests on the formal conception of algorithms and decidability, when applied directly, tell us that all objects in the world are computers, since they are physically implementing computable functions. Put another way, if you wanted to calculate a parabolic curve you could throw a ball and simply track its movement, so in some sense, you could use the ball to solve your mathematical problem, and it is thus a "computer" solving your parabolic curve. Though this is formally correct it is conceptually unsatisfying. What use is it for us to define "computer" in this manner if it trivially renders most of the universe and everything within it a computer?

In this instance, the *use* of the word becomes critical. Despite the formal definitions, computer scientists rarely refer to thrown balls as "computers". Is that because computer scientists only use the word to refer to electronic devices like our laptops and smartphones? No, there are clear examples of references in computer science to computers that are very different from the typical digital computers we're all familiar with, including analog computers, quantum computers, stochastic computers, DNA computers, and neuromorphic computers (Gaines, 1967; Rubel, 1993; Adleman, 1994, 1998; Beaver, 1995; Paun et al., 2005; Van Noort and Landweber, 2005; Elbaz et al., 2010; Ladd et al., 2010; Furber, 2016; Schuman et al., 2017; Tsividis, 2018; van de Burgt et al., 2018; Shastri et al., 2021). None of these forms of computer operate like a laptop or smartphone; they can use analog signals, stochastic operations, parallel calculations, biological substrates, etc. And yet, the usage of the word "computer" in such articles does not appear to be intended as a metaphor. So, what then renders something a "computer" in computer science, according to the way the word is used?

What we can see in research papers is that computer scientists generally use the word "computer" to refer to any physical machinery that can, *in theory*, implement *any* computable function (per the definition above), i.e., a physical system that in principle can serve as a "universal" computation device (Beaver, 1995; Van Noort and Landweber, 2005; Ladd et al., 2010). For example, when Adleman (1994) closed his paper on DNA-based computation he said, "*One can imagine the eventual emergence of a general purpose computer consisting of nothing more than a single macromolecule conjugated to ribosome like collection of enzymes that act on it*". Here, the key point is the words "general purpose". It is the potential for general purpose computation with DNA that, we argue, makes computer scientists inclined to talk about "DNA computers", despite the fact that a macromolecule conjugated to a ribsome like collection of enzymes would engage in calculation in a very different manner than a modern silicon chip.

Note also our use of the phrase "*in theory*", above. Many of the systems that computer scientists refer to as "computers" cannot in practice implement any computable function due to size, memory, time, noise, and energy limitations. So, for example, quantum computers are not yet capable of computing any computable function, but in theory they could, and so we refer to them as "computers". And, of course, a laptop is a "computer" because it can be shown that the operations it utilizes could theoretically implement any computable function, though in reality some functions would take too long or require too much memory (e.g., calculating the number of prime numbers less than $10^{10^{10^{10}}}$). In contrast, a thrown ball is limited to implementing only those functions that describe its movement through space. Thus, when computer scientists use the word "computer", they generally use it to refer only to physical machinery that could, in theory, compute any computable function, which is by no means applicable to most things (Adleman, 1998; Elbaz et al., 2010; Ladd et al., 2010; van de Burgt et al., 2018).

3.3. Applying the Definition From Computer Science to Brains

Given the use-based definition above (physical machinery that can implement any computable function in theory), are brains computers? The answer for most scientists should be *yes*. First, though there is disagreement in philosophy as to whether brains are purely physical systems and whether their operations rely solely on physical machinery, the perspective of physicalism is widely accepted by brain scientists and we are not aware of any brain scientists who doubt that the operations of the brain are fundamentally physical. Second, with the aid of a pencil and paper, a human brain can in theory implement any program that one could implement with modern digital computers. The only limits would be time and energy, which as noted, also apply to other computers, like laptops. Even without pencil and paper, the only real limit to a person implementing any computer program is again the limits on their memory, time, and energy, not their general capabilities, *per se*. Conceptually, we can perform all of the same operations specified by the languages that we program our laptops with. Third, and perhaps more importantly, if one is concerned with practical implications for the brain sciences, real neural circuits are in theory, likely capable of implementing all of the functions that artificial neural networks (ANNs) can, if not more. And, computer scientists have shown that ANNs can implement any computable function (Hornik, 1991; Siegelmann and Sontag, 1995). In other words, as long as real brains have the same or greater capabilities than ANNs (again ignoring memory, time, and energy constraints), then they are surely capable of implementing any computable function.

Therefore, according to the use-based definition of "computer" in computer science, brains are literally computers. *There is no metaphor*. The claim here is not that brains work anything like our laptops and smartphones. But the use-based definition of "computer" in computer science isn't "something that works like a laptop or smartphone"—DNA and quantum bits are very different from silicon chips. The definition of "computer" is physical machinery that, in theory, can implement any computable function, and brains meet this definition at least as well as many of the other devices that we all refer to as "computers" on a regular basis with no complaint and no hint of a metaphor.

We should address here a few of the common misconceptions that lead people to object to this line of logic. First, one of the most common points of confusion is that some people think that the formal definition based on the Church-Turing thesis implies that to be a computer an object's internal machinery itself must operate in a similar manner to Turing machines (Fodor, 1981; Copeland, 2020). But, this is simply a misunderstanding, as many types of computers (e.g., analog computers or neuromorphic computers) do not operate like a Turing machine. This misunderstanding may derive from the fact that modern digital computers bear some resemblance to the Turing machine formalism. But importantly, Turing machines are just mathematical constructs—they are sets of rules, not physical machines. Your laptop computer is no more a Turing machine than it is a lambda calculus. Nothing about the way computer scientists use the word "computer" demands that the object work like a Turing machine—the object in question must simply be capable of implementing the same functions as Turing machines.

Second, another reasonably common claim is that brains can't be computers because they can solve problems that are not decideable (Penrose, 1989; Siegelmann, 1995). We note that no one has ever convincingly demonstrated that brains can actually do this. However, importantly, this claim speaks to the question of whether brains are literally computers, not whether computers are a good metaphor for brains. As such, though it is an interesting objection that warrants consideration, it does not change our fundamental point, which is that there is no metaphor in play when we apply the definition of the word "computer" from computer science to brains. Finally, another source of confusion can enter into the discussion when simulations are discussed. We can, of course, simulate aspects of how neural circuits work using digital computers. And so, it has sometimes been believed that the claim that brains are computers derives from our ability to simulate them, and in turn, it has been (rightly) pointed out that the ability to simulate something with a computer does not make that thing a computer (Brette, 2018), e.g., we can simulate a ball bouncing but that does not make a ball a computer. But, as outlined above, it is not our ability to simulate neural circuits that makes brains computers, it is their theoretical ability to implement any computable function. Hence, the question of simulation is actually irrelevant to the question of whether brains are computers or not. The only relevant question is: Can brains implement any computable function in theory? And we argue that the answer is certainly "yes".

4. THE USE OF THE WORD "COMPUTER" OUTSIDE OF COMPUTER SCIENCE

All of this may be a bit surprising to many readers, because the definitions of "computer" given above is not how the average person, nor the average academic outside of computer science, understands and uses this word. As such, we may ask for an alternative definition of "computer", one that aligns better with the usage of people outside of computer science.

When most people speak of a "computer" today, they use the word to refer to human-made electronic devices that can perform complex mathematical calculations, display multimedia content, and communicate with other similar devices. According to this usage, a computer can be defined as something like "an electronic appliance that we can use for calculation, communication, and entertainment". Obviously, this definition does not apply to brains, nor would it serve as a particularly good metaphor either.

Within academia, there are also people in the brain sciences and philosophy who are more knowledgeable about computers (and brains) but who are still only partially familiar with the ideas from computer science presented above. For these people, the usage of the word "computer" often still centers on the human-made electronic devices we are all familiar with, but it includes some more details of how those devices work. Specifically, the vast majority of modern digital computers are extensions of the "Von Neumann architecture", first developed by the polymath John Von Neumann in the 1940's (Von Neumann, 1993). Though there have been changes to Von Neumann's original design (Godfrey and Hendry, 1993), some of his ideas are still central to modern digital computers. These include the use of a central processing unit (CPU) for sequential operations of arithmetic logic, a control unit in the CPU that stores the sequence of instructions for the CPU to perform, a random access memory (RAM) module for storing intermediate calculations, and an external memory (or "hard drive") for long-term storage of information. It's interesting to note that Von Neumann's designs are reminiscent of how we define Turing machines, with an internal state, and a step-by-step processing of input symbols to produce output. Given this apparent similarity, many writers use the word "computer" to mean something like "human-made machines that have the qualities of Von Neumann architecture machines, and which resemble aspects of Turing machines" (Cisek, 1999; Epstein, 2016; Cobb, 2020). Hence, one can find articles where people refer to computers and computation as being necessarily sequential, or discrete, or restricted to passive processing of a stream of inputs using a step-by-step program (Van Gelder, 1998; Cisek, 1999; Brette, 2018, 2019; Cobb, 2020). For example, Cisek (1999) notes the importance of control for brains and animals, which he argues is ignored by the computer metaphor for the brain, because it instead presupposes that "*...perception is like input, action is like output, and all the things in-between are like the information processing performed by computers.*" His point here is that brains are not simply taking inputs and producing outputs based on some internal state (akin to the formalism of Turing machines), but rather, they are constantly engaged in adaptive interactions for controlling the body and the world in order to achieve specific ends. However, control is something that people in computer science would happily say computers can do (Arnăutu and Neittaanmäki, 2003). Thus, Cisek (1999)'s concern is less about "computers" as they are defined in computer science, and more about "computers" as they are defined by those outside of computer science.

With the definition from outside of computer science in hand, are brains computers? Most certainly not. Brains do not use sequential processing—quite the opposite they use massively parallel processing (Rumelhart et al., 1988). Brains do not use discrete symbols stored in memory registers—they operate on high-dimensional, distributed representations stored *via* complex and incompletely understood biophysical dynamics (Jazayeri and Ostojic, 2021). And, brains do not passively process inputs to generate outputs using a step-by-step program—they control an embodied, active agent that is continuously interacting with and modulating the very systems that generate the sensory data they receive in order to achieve certain goals (Cisek, 1999; Brette, 2019). Thus, with the definition from outside of computer science we can say not only that brains are not computers, we can also say that computers are poor metaphors for brains, since the manner in which they operate is radically different from how brains operate.

There are some complications to this that should be noted. First, brains are capable of some forms of more traditional tasks that our digital computers are good at, i.e., various forms of discrete, sequential processing (Fodor, 1981; Marcus, 2015). For example, people can do long-division, symbolic logic, list sorting, etc. So, we might say that computers (according to the definition from outside of computer science) can serve as reasonable metaphors for *some types* of human cognition. Moreover, modern digital computers are rapidly evolving to incorporate more parallel, distributed, dynamic operations (Shukur et al., 2020), and some engineers are actively trying to explicitly mimic the operations of brains using "neuromorphic" chips (Furber, 2016; Schuman et al., 2017; van de Burgt et al., 2018; Shastri et al., 2021). These more modern forms of human-made computers present some complications for the use-based definition of "computer" from outside of computer science. Nonetheless, if we are committed to the concept of use-based meaning, then we can say that when some authors dismiss the brain-computer metaphor (Carello et al., 1984; Cisek, 1999; Brette, 2018) they are using the word "computer" to mean something more like traditional, Von Neumann architecture machines, not neuromorphic chips, etc. And, as noted, such authors are correct, brains are not very much like these traditional digital computers.

5. DISCUSSION

Tying our two different threads together, we can conclude that the question of whether brains are computers (or like computers) is really a matter of semantics: it depends on which definition you are using. If you adopt the definition of "computer" based on how computer scientists use the word (to refer to physical machinery that can theoretically engage in any decidable computation), then brains are literally computers. Alternatively, if we adopt the definition of "computer" based on the usage from outside of computer science (to refer to devices that sequentially and discretely process inputs in a passive manner), then brains are not computers, and at best, computers serve as a weak metaphor for only a limited slice of human cognition. The message that we are providing here to the brain sciences community is, we hope, very clear: brains are either literally computers, or really not much like computers, depending on the definition we employ. Thus, it is ultimately a matter of semantics, and arguably, debates about the "brain-computer metaphor" are not productive. We can simply stop engaging in them.

It is worth noting that our argument here rests on an important stance vis-à-vis the philosophy of science. Specifically, we are assuming that scientists can and do use words and concepts in a literal manner. This is in contrast to a potential perspective that views all concepts as metaphors (Lakoff and Johnson, 1980). Putting aside the larger philosophical debate that would be possible on this matter, we wish here simply to clarify and recognize that our perspective very much so rests on the idea that there are non-metaphorical uses of words and concepts in science.

The natural question that emerges from the realization that the brain-computer metaphor debate is actually just a semantic disagreement is to ask whether it matters which definition of "computer" we adopt? Does it affect the brain sciences in any meaningful way to adopt one definition or the other? In particular, should the field be concerned with the definition from computer science at all, given that it is not terribly intuitive and not what most people in the brain sciences think of when they hear the word "computer"?

We would argue that the definition we adopt is very important, and both definitions should be considered. The usage of "computer" in computer science can actually be very useful for the brain sciences in some circumstances. The reason is that when one realizes that brains are literally computers (in the computer science sense of the word) then much of the theory about computation from computer science is applicable to brains. This connection is what opens up space in computational neuroscience to explore the brain using conceptual tools from computer science and AI, which has produced both important insights in neuroscience (Richards et al., 2019) and advances in AI (Hassabis et al., 2017). Indeed, asking the question, "*What sort of computer is the brain?*", is arguably the underpinning of modern neural networks (Rumelhart et al., 1988), which have been very useful for the brain sciences. Asking this question is how we arrive at core concepts in computational neuroscience such as parallel processing, content addressable memory, and spike-based computation. Similarly, consider the question of randomness in computation. Thanks to our understanding that the brain is a computer we can apply concepts from computer science, such as convergence and constraint satisfaction, to better understand the normative importance of stochastic vesicle release in neurons (Maass and Zador, 1999; Habenschuss et al., 2013). Similarly, concepts from compression theory help us to understand the nature of representations in the brain (Olshausen and Field, 1996) and dynamic programming concepts used in reinforcement learning help us to understand memory replay (Mattar and Daw, 2018). More broadly, the inter-disciplinary intersection between AI and the brain sciences depends on the computer science definition of the word "computer", and so, if we reject this definition outright we risk shutting the door on a very active field of research that has proven very fruitful for both the brain sciences and AI. At the same time, it is worth being vigilant and clear that brains do not work like our laptops and smartphones, and these devices serve as a poor metaphor for brains. So, depending on the audience and the purpose of the work, sometimes we should adopt the definition from outside of computer science, as long as we are clear on what that definition of "computer" actually implies. There is no single correct definition for "computer"—but we all must be clear on what we mean when we write and speak. On this point, the vast majority of researchers across all disciplines must surely agree.

AUTHOR CONTRIBUTIONS

BR and TL co-wrote the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by NSERC (Discovery Grant: RGPIN-2020-05105; Discovery Accelerator Supplement: RGPAS-2020-00031), CIFAR (Canada CIFAR AI Chair and Learning in Machines and Brains Program), and Healthy Brains, Healthy Lives (New Investigator Award: 2b-NISU-8).

REFERENCES

Adleman, L. M. (1994). Molecular computation of solutions to combinatorial problems. *Science* 266, 1021–1024. doi: 10.1126/science.7973651

Adleman, L. M. (1998). Computing with DNA. *Sci. Am*. 279, 54–61. doi: 10.1038/scientificamerican0898-54

Arnǎutu, V., and Neittaanmäki, P. (2003). *Optimal Control from Theory to Computer Programs*. Berlin: Springer Science & Business Media. doi: 10.1007/978-94-017-2488-3

Beaver, D. (1995). A universal molecular computer. *DNA Based Comput*. 27, 29–36. doi: 10.1090/dimacs/027/03

Brette, R. (2018). *What Is Computational Neuroscience? Is the Brain a Computer*? Available online at: http://romainbrette.fr/what-is-computational-neuroscience-xxx-is-the-brain-a-computer/

Brette, R. (2019). Is coding a relevant metaphor for the brain? *Behav. Brain Sci*. 42, e243. doi: 10.1017/S0140525X19001997

Carello, C., Turvey, M. T., Kugler, P. N., and Shaw, R. E. (1984). "Inadequacies of the computer metaphor," in *Handbook of Cognitive Neuroscience*, ed M. S. Gazzaniga (Boston, MA: Springer), 229–248. doi: 10.1007/978-1-4899-2177-2_12

Church, A. (1936a). A note on the Entscheidungsproblem. *J. Symbol. Logic* 1, 40–41. doi: 10.2307/2269326

Church, A. (1936b). An unsolvable problem of elementary number theory. *Am. J. Math*. 58, 345–363. doi: 10.2307/2371045

Cisek, P. (1999). Beyond the computer metaphor: behaviour as interaction. *J. Conscious. Stud*. 6, 125–142.

Cobb, M. (2020). *Why Your Brain Is Not a Computer*. The Guardian. London. Available online at: https://www.theguardian.com/science/2020/feb/27/why-your-brain-is-not-a-computer-neuroscience-neural-networks-consciousness (accessed February 27, 2020).

Cook, S. A. (1992). "Computability and complexity of higher type functions," in *Logic from Computer Science*, ed Y. N. Moschovakis (New York, NY: Springer), 51–72. doi: 10.1007/978-1-4612-2822-6_3

Cook, S. A. (2014). Conversations: from Alan Turing to NP-completeness. *Curr. Sci*. 106, 1696. doi: 10.18520/cs/v106/i12/1696-1701

Copeland, B. J. (1997). The broad conception of computation. *Am. Behav. Sci*. 40, 690–716. doi: 10.1177/0002764297040006003

Copeland, B. J. (2020). "The Church-Turing thesis," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta (Metaphysics Research Lab, Stanford University).

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms*. Cambridge, MA: MIT Press.

Diamant, E. (2008). Unveiling the mystery of visual information processing in human brain. *Brain Res*. 1225, 171–178. doi: 10.1016/j.brainres.2008.05.017

Dreyfus, H. L., and Hubert, L. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press.

Elbaz, J., Lioubashevski, O., Wang, F., Remacle, F., Levine, R. D., and Willner, I. (2010). DNA computing circuits using libraries of DNAzyme subunits. *Nat. Nanotechnol*. 5, 417–422. doi: 10.1038/nnano.2010.88

Epstein, R. (2016). *Your Brain Does Not Process Information and It Is Not a Computer*. Available online at: https://aeon.co/essays/your-brain-does-not-process-information-and-it-is-not-a-computer

Fodor, J. A. (1981). The mind-body problem. *Sci. Am*. 244, 114–123. doi: 10.1038/scientificamerican0181-114

Furber, S. (2016). Large-scale neuromorphic computing systems. *J. Neural Eng*. 13, 051001. doi: 10.1088/1741-2560/13/5/051001

Gaines, B. R. (1967). Stochastic computing, in *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference AFIPS '67 Spring* (New York, NY: Association for Computing Machinery), 149–156. doi: 10.1145/1465482.1465505

Godfrey, M. D., and Hendry, D. F. (1993). The computer as von Neumann planned it. *IEEE Ann. History Comput*. 15, 11–21. doi: 10.1109/85.194088

Grier, D. A. (2001). Human computers: the first pioneers of the information age. *Endeavour* 25, 28–32. doi: 10.1016/s0160-9327(00)01338-7

Habenschuss, S., Jonke, Z., and Maass, W. (2013). Stochastic computations in cortical microcircuit models. *PLoS Comput. Biol*. 9, e1003311. doi: 10.1371/journal.pcbi.1003311

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron* 95, 245–258. doi: 10.1016/j.neuron.2017.06.011

Hilbert, D. (1902). Mathematical problems. *Bull. Am. Math. Soc*. 8, 437–479. doi: 10.1090/S0002-9904-1902-00923-3

Hilbert, D., and Ackermann, W. (1999). *Principles of Mathematical Logic*, Vol. 69. Providence, RI: American Mathematical Society.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Netw*. 4, 251–257. doi: 10.1016/0893-6080(91)90009-T

Hunt, E. (1989). Cognitive science: definition, status, and questions. *Annu. Rev. Psychol*. 40, 603–629. doi: 10.1146/annurev.ps.40.020189.003131

Jazayeri, M., and Ostojic, S. (2021). Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Curr. Opin. Neurobiol*. 70, 113–120. doi: 10.1016/j.conb.2021.08.002

Kwisthout, J., and van Rooij, I. (2020). Computational resource demands of a predictive Bayesian brain. *Comput. Brain Behav*. 3, 174–188. doi: 10.1007/s42113-019-00032-3

Ladd, T. D., Jelezko, F., Laflamme, R., Nakamura, Y., Monroe, C., and O'Brien, J. L. (2010). Quantum computers. *Nature* 464, 45–53. doi: 10.1038/nature08812

Lakoff, G., and Johnson, M. (1980). *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

Maass, W. (2016). Searching for principles of brain computation. *Curr. Opin. Behav. Sci*. 11, 81–92. doi: 10.1016/j.cobeha.2016.06.003

Maass, W., and Zador, A. M. (1999). Dynamic stochastic synapses as computational units. *Neural Comput*. 11, 903–917. doi: 10.1162/089976699300016494

Maccormac, E. R. (1986). "Men and machines: the computational metaphor," in *Philosophy and Technology II* (ed C. Mitcham and A. Huning (Berlin: Springer), 157–170. doi: 10.1007/978-94-009-4512-8_11

Marcus, G. (2015). *Opinion - Face It, Your Brain Is a Computer*. New York, NY: The New York Times.

Mattar, M. G., and Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci*. 21, 1609–1617. doi: 10.1038/s41593-018-0232-z

Newell, A. (1980). Physical symbol systems. *Cogn. Sci*. 4, 135–183. doi: 10.1207/s15516709cog0402_2

Olshausen, B. A., and Field, D. J. (1996). Natural image statistics and efficient coding. *Netw. Comput. Neural Syst*. 7, 333. doi: 10.1088/0954-898X_7_2_014

Oteiza, P., Odstrcil, I., Lauder, G., Portugues, R., and Engert, F. (2017). A novel mechanism for mechanosensory-based rheotaxis in larval zebrafish. *Nature* 547, 445–448. doi: 10.1038/nature23014

Paun, G., Rozenberg, G., and Salomaa, A. (2005). *DNA Computing: New Computing Paradigms*. Berlin: Springer Science & Business Media.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. New York, NY: Oxford University Press. doi: 10.1093/oso/9780198519737.001.0001

Pylyshyn, Z. W. (1984). *Computation and Cognition: Towards a Foundation for Cognitive Science*. Cambridge, MA: MIT Press.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., et al. (2019). A deep learning framework for neuroscience. *Nat. Neurosci*. 22, 1761–1770. doi: 10.1038/s41593-019-0520-2

Rubel, L. A. (1993). The extended analog computer. *Adv. Appl. Math*. 14, 39–50. doi: 10.1006/aama.1993.1003

Rumelhart, D. E., McClelland, J. L., and PDP Research Group (1988). *Parallel Distributed Processing*, Vol.1. Cambridge: MIT Press. doi: 10.1016/B978-1-4832-1446-7.50010-8

Schuman, C. D., Potok, T. E., Patton, R. M., Birdwell, J. D., Dean, M. E., Rose, G. S., et al. (2017). A survey of neuromorphic computing and neural networks in hardware. *arXiv [Preprint] arXiv*:1705.06963.

Shastri, B. J., Tait, A. N., Ferreira de Lima, T., Pernice, W. H. P., Bhaskaran, H., Wright, C. D., et al. (2021). Photonics for artificial intelligence and neuromorphic computing. *Nat. Photon*. 15, 102–114. doi: 10.1038/s41566-020-00754-y

Shukur, H., Zeebaree, S. R., Ahmed, A. J., Zebari, R. R., Ahmed, O., Tahir, B. S. A., et al. (2020). A state of art survey for concurrent computation and clustering of parallel computing for distributed systems. *J. Appl. Sci. Technol. Trends* 1, 148–154. doi: 10.38094/jastt1466

Siegelmann, H., and Sontag, E. (1995). On the computational power of neural nets. *J. Comput. Syst. Sci*. 50, 132–150. doi: 10.1006/jcss.1995.1013

Siegelmann, H. T. (1995). Computation beyond the turing limit. *Science* 268, 545–548. doi: 10.1126/science.268.5210.545

Simon, H. A. (1980). Cognitive science: the newest science of the artificial. *Cogn. Sci*. 4, 33–46. doi: 10.1207/s15516709cog0401_2

Smith, C. U. (1993). The use and abuse of metaphors in the history of brain science. *J. History Neurosci*. 2, 283–301. doi: 10.1080/09647049309525577

Tsividis, Y. (2018). Not your Father's analog computer. *IEEE Spectrum* 55, 38–43. doi: 10.1109/MSPEC.2018.8278135

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *J. Math*. 58, 5.

Turing, A. M. (1937). Computability and λ-definability. *J. Symbol. Logic* 2, 153–163. doi: 10.2307/2268280

van de Burgt, Y., Melianas, A., Keene, S. T., Malliaras, G., and Salleo, A. (2018). Organic electronics for neuromorphic computing. *Nat. Electron*. 1, 386–397. doi: 10.1038/s41928-018-0103-3

Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behav. Brain Sci*. 21, 615–628. doi: 10.1017/S0140525X98001733

Van Noort, D., and Landweber, L. F. (2005). Towards a re-programmable DNA computer. *Nat. Comput*. 4, 163–175. doi: 10.1007/s11047-004-4010-3

Vlasits, A. (2017). *Tech Metaphors are Holding Back Brain Research*. Wired. (San Francisco, CA). Available online at: https://www.wired.com/story/tech-metaphors-are-holding-back-brain-research/ (accessed June 12, 2017).

Von Neumann, J. (1993). First draft of a report on the EDVAC. *IEEE Ann. History Comput*. 15, 27–75. doi: 10.1109/85.238389

West, D. M., and Travis, L. E. (1991). The computational metaphor and artificial intelligence: a reflective examination of a theoretical falsework. *AI magazine* 12, 64–64.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Basil Blackwell. Available online at: https://static1.squarespace.com/static/54889e73e4b0a2c1f9891289/t/564b61a4e4b04eca59c4d232/1447780772744/Ludwig.Wittgenstein.-.Philosophical.Investigations.pdf

Conflict of Interest: TL is employed by DeepMind Inc., and BR works as an external consultant for DeepMind Inc.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 20 December 2021
doi: 10.3389/fevo.2021.802300

Why Can the Brain (and Not a Computer) Make Sense of the Liar Paradox?

Patrick Fraser [1]*, *Ricard Solé* [2,3] *and Gemma De las Cuevas* [4]

[1] Department of Philosophy, University of Toronto, Toronto, ON, Canada, [2] ICREA Complex Systems Lab, University Pompeu Fabra, Barcelona, Catalonia, [3] Santa Fe Institute, Santa Fe, NM, United States, [4] Institute for Theoretical Physics, Innsbruck, Austria

Ordinary computing machines prohibit self-reference because it leads to logical inconsistencies and undecidability. In contrast, the human mind can understand self-referential statements without necessitating physically impossible brain states. Why can the brain make sense of self-reference? Here, we address this question by defining the Strange Loop Model, which features causal feedback between two brain modules, and circumvents the paradoxes of self-reference and negation by unfolding the inconsistency in time. We also argue that the metastable dynamics of the brain inhibit and terminate unhalting inferences. Finally, we show that the representation of logical inconsistencies in the Strange Loop Model leads to causal incongruence between brain subsystems in Integrated Information Theory.

Keywords: self-reference, cognition, consciousness, computation, causal structure, integrated information theory

OPEN ACCESS

Edited by:
Pedro Martinez,
University of Barcelona, Spain

Reviewed by:
Eric Smith,
The University of Texas at El Paso, United States
George Kampis,
Eötvös Loránd University, Hungary

****Correspondence:***
Patrick Fraser
p.fraser@mail.utoronto.ca

Specialty section:
This article was submitted to Models in Ecology and Evolution, a section of the journal Frontiers in Ecology and Evolution

Received: *26 October 2021*
Accepted: *25 November 2021*
Published: *20 December 2021*

Citation:
Fraser P, Solé R and De las Cuevas G (2021) Why Can the Brain (and Not a Computer) Make Sense of the Liar Paradox? Front. Ecol. Evol. 9:802300. doi: 10.3389/fevo.2021.802300

1. INTRODUCTION

Are brains like computers? Can technological metaphors provide satisfactory explanations for the complexity of human brains (and brains in general)? Before electronic computers became a reality, some versions of the previous questions had always been there. In the seventeenth century, the development of mechanical clocks and later on mechanical automata led to questions with far-reaching philosophical implications, such as the possibility of creating a mechanical human and an artificial mind (by René Descartes and others Wood, 2002). Later, brains and machines were compared to electric batteries (since it became clear that electricity was involved in brain processes), and early works by visionaries such as Alfred Smee represented brains and the activity of thinking in terms of networks of connected batteries (Smee, 1850). Other network-level metaphors of the brain such as telegraphs and telephone webs replaced the old ones, until the metaphor of the computer prevailed in the 1950s (Cobb, 2020).

The computer was apparently the right metaphor: It could store large amounts of data, manipulate them and perform complex input-output tasks that involved information processing. Additionally, the new wave of computing machines provided an appropriate technological context to simulate logical elements similar to those present in nervous systems. Theoretical developments within mathematical biology by McCulloch and Pitts (1943) revealed one first major result: The units of cognition—neurons—could be described with a formal framework. Formal neurons were described in terms of threshold units, largely inspired by the state-of-the-art knowledge of real neurons (Rashevsky, 1960). Over the last decades, major quantitative advances have been obtained by combining neuron-inspired models with multilayer architecture (LeCun et al., 2015) and physics of neuromorphic computing (Indiveri and Liu, 2015; Markovi et al., 2020). These developments are largely grounded in early theories (Rumelhart et al., 1986; Fukushima, 1988) with novel hardware improvements and a massive use of training data.

Despite the obvious success of computing and information technology, we are still far from the dream of building or simulating a truly intelligent system. To begin with, computers and their abstract representation in terms of Turing machines are highly modular, programmable and sequential (Arbib, 2012) (see **Figure 1**). Instead, neural systems are the result of evolutionary tinkering and selection that favored exploiting redundancy and parallelism (Allman, 1999; Martinez and Sprecher, 2020). That does not prohibit the existence of interesting links that help make sense of brain in terms of Turing machines: Many functional responses of brains are essentially sequential in nature, despite the highly parallel integration that feeds serial (and slow) cognitive task production (Zylberberg et al., 2011). Yet, the most remarkable departure of brains from computers is probably the presence of re-entrant circuits, i.e., the recursive exchange of signals across multiple, parallel and reciprocal connections (Edelman, 1992). Indeed, some authors have posited that closed feedback loops are crucial for conscious experience (Hofstadter, 1979; Oizumi et al., 2014). Are closed feedback loops the key for a formal differentiation between brains and computers? Closed feedback loops can allow for self-reference (Grim, 1993), and the human brain is capable of self-referential inference. So this begs the question: Why can the brain make sense of self-reference, whereas a computer can't?

We address this question by considering paradoxes of self-reference and negation (Prokopenko et al., 2019). Studies in logic, linguistics, and general philosophy for many centuries have illustrated that when statements negatively refer to their own features, contradictions follow in short order. This is made clear from sentences such as:

The sentence presently being uttered is false. (1)

Taking this sentence at its word—supposing it to be true—we find out it is false. However, taking it to be false, we are forced to

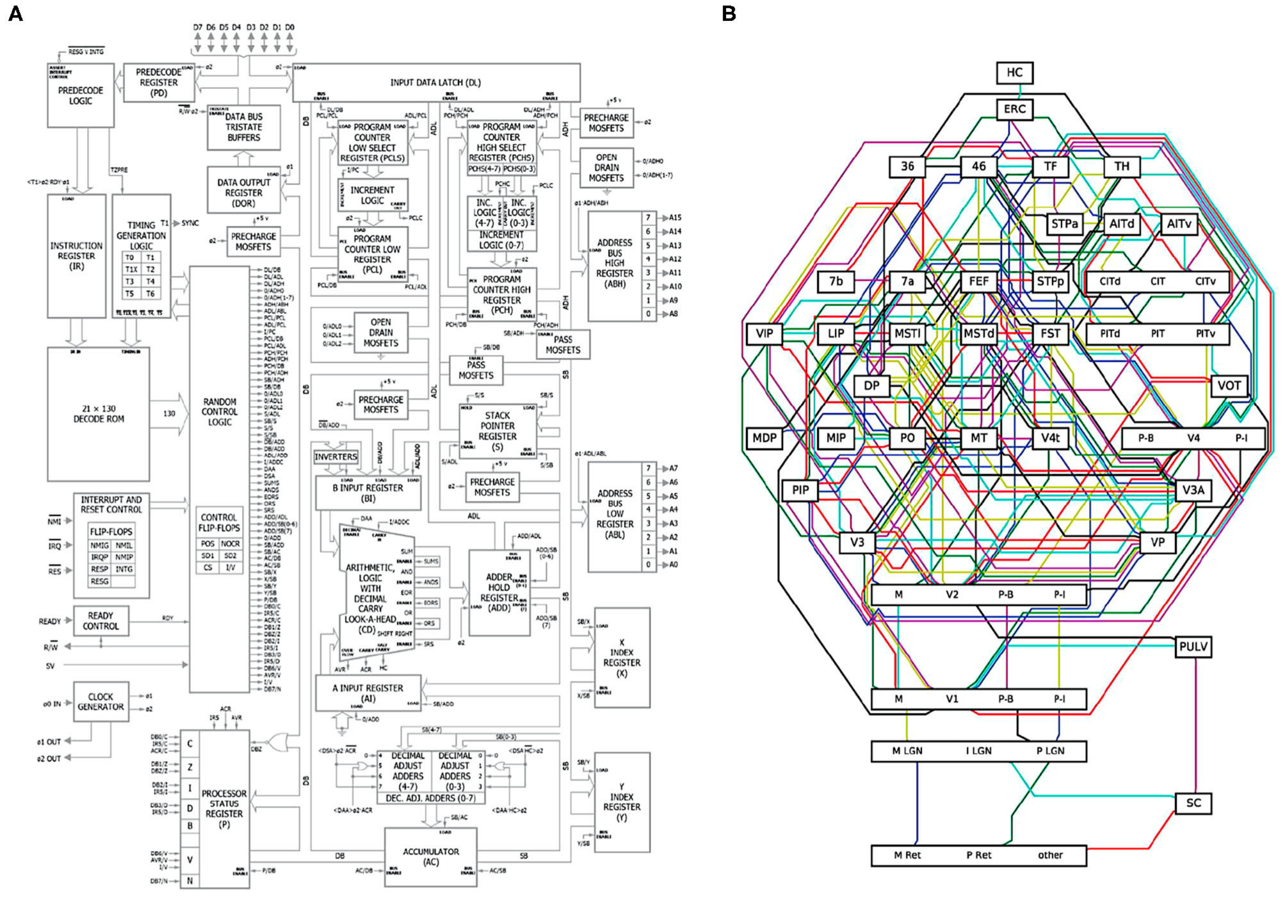


FIGURE 1 | Computer vs. brain architecture. A topological analysis of **(A)** computer chips and **(B)** brains (visual cortex organization) reveals fundamental dissimilarities. These include the strict modular organization of the former contrasted with the highly parallel, integrated architecture of the latter. The circuits responsible for higher-order cognitive brain tasks display re-entrant feedback loops that are absent on the *in-silico* counterparts (compare with **Figure 2**). Image adapted from Jonas and Kording (2017).

conclude that it is true. When we assign truth values to sentences, we classically assume that truth and falsity are mutually exclusive and exhaustive, yet self-referential sentences appear to have *over-determined* truth values (Priest, 2006, pp. 14–15): We are obliged to evaluate them simultaneously as true and false, a contradiction. This may compel the logician to use formal languages that block such self-referential constructions to preserve their consistency at the cost of limiting their expressiveness. Such a pursuit of consistency is perhaps well-motivated in purely formal settings such as mathematics, but self-reference is readily available within natural language, and human minds are capable of formulating and thinking about self-referential paradoxes and becoming aware of their inconsistency.

Computers are incapable of resolving a paradox such as sentence 1—they get caught in endless loops—, whereas the brain can "reason" about this paradox. Let us examine the latter statement by bringing forward some basic facts about the workings of the brain. In the ordinary course of experience, our state of mind may possess many subtle and composite features, but we only ever occupy one such mental state at a time: There are no "superpositions" of mental states. Furthermore, if we take mental states to be somehow derivative of brain states (by whatever account of the emergence of consciousness one prefers), the deterministic or unitary evolution of physical systems given by our best physical theories suggests that our brains only ever occupy a single physical state.[1] Whatever the mechanism responsible for the emergence of mental states from brain states is, surely the brain state that grounds the awareness of some fact is different from the brain state that grounds the awareness of its negation. Thus, occupying a mental state corresponding to awareness of a contradiction would *seem* to be a physical impossibility *par excellence* insofar as it would necessitate one's brain to be in two distinct states at once. Yet, upon interpreting the sentence 1, the reader comes to think about a self-referential statement and understand its contradictory nature, and so the cognitive processing of self-referential statements is clearly *not* a physical impossibility (nor do they get stuck in an unhalting cycle of thoughts one might expect of a machine tasked with deciding the truth value of such a sentence). How is this possible?

In this paper, we address this question by constructing a high-level model of the brain, termed a Strange Loop Model (Section 2), from which we conclude that:

1. The brain makes sense of self-reference by spreading out inconsistent truth values in time, thereby avoiding physically impossible states (Section 3.1).
2. The representation of logical inconsistencies in the brain leads to causal incongruence between brain subsystems (Section 3.3).
3. The metastable dynamics of the brain and its interactions with external stimuli inhibit and terminate unhalting inferences (Section 3.4).

Statement 1 says that the brain represents and processes self-referential sentences by treating their truth values as dynamical quantities. It follows that the resulting contradictions are unfolded in time, and thus do not require physically impossible brain states. Statement 2 describes how this "unfolding" works: Different parts of the brain yield disagreeing predictions about the brain's future states, and this disagreement is made apparent by analyzing the causal feedback between these parts. This disagreement is known in Integrated Information Theory (IIT) as *incongruence* (Albantakis and Tononi, 2019). This causal feedback is not encountered in Turing machines because they are feed-forward systems. Statement 3 claims that the brain does not succumb to halting problems when processing statements whose truth values are undecidable, because the metastable nature of brain dynamics precludes falling into lock-in states (Tognoli and Kelso, 2014).

This paper is structured as follows. We present the Strange Loop Model (SLM) of the brain (Section 2), and we use it to represent self-referential inferences in the brain (Section 3). Finally we conclude and discuss further directions (Section 4).

2. THE STRANGE LOOP MODEL

Here we present a high-level model of the brain by describing it as a discrete dynamical system (Section 2.1), partitioning it into functionally distinct modules (Section 2.2), and investigating their causal structure (Section 2.3). The name originates from Hofstadter (1979, 2007): Strange loops arise when, by moving only upwards (or downwards) in a hierarchy, one encounters oneself at the same place where one started.

2.1. Discrete Dynamics of Brain Modules

Here we describe the brain as a discrete dynamical network of connectomic units (Sporns et al., 2005). We consider that n such units (indexed $i = 1, \dots, n$), evolving in discrete time $t \in \mathbb{Z}$, and denote the state of unit i at time t by $x_i^t \in \Sigma_i$, where Σ_i is a finite state space. The state of the "brain" in the SLM at time t is denoted

$$B^t = (x_1^t, \dots, x_n^t) \in \Sigma_1 \times \dots \times \Sigma_n =: \Sigma.$$

The dynamics of such a system are given by a transition function $\mathcal{T} : \Sigma \to \Sigma$ so that $B^{t+1} = \mathcal{T}(B^t)$ and we denote ith component of $\mathcal{T}(B^t)$ by $\mathcal{T}_i(B^t) := x_i^{t+1}$.

We consider a probability distribution p on Σ. For any $z \in \Sigma_i$, the conditional probability (also denoted p) is defined as

$$p(z|B^t) = \begin{cases} 1 & \text{if } z = \mathcal{T}_i(B^t) \\ 0 & \text{else.} \end{cases}$$

We suppose that all units are conditionally independent at any given time $t \in \mathbb{Z}$, so they satisfy:

$$p(B^{t+1}|B^t) = \prod_{i=1}^{n} p(x_i^{t+1}|B^t). \tag{2}$$

Additionally, we suppose that the future state of the brain depends only on the immediately preceding state (Markovianity),

[1] Since the brain is fundamentally a quantum system, this physical state *could* be in a superposition, but as Tegmark (2000) has shown, even neurons are sufficiently macroscopic systems that decoherence would likely prevent quantum effects from being relevant.

so that if $t_1 < t_2 < \cdots < T$, the joint probability distribution factors as

$$p(B^{t_1}, B^{t_2}, \ldots, B^T) = p(B^{t_1}) \prod_{n=1}^{T-1} p(B^{t_{n+1}} | B^{t_n}). \tag{3}$$

With this setup one may use the intervention calculus from probabilistic causal modeling (e.g., as elaborated by Pearl, 2009) to understand how connectomic units causally influence each other. Following the exposition in Krohn and Ostwald (2017), given any two subsystems $X, Y \subseteq B$, one defines the *effect probability* p_e, the joint *cause-effect probability* p_{ce}, and the *cause probability* p_c to be:

$$\begin{aligned} p_e(Y^t | X^{t-1}) &:= p(Y^t | X^{t-1}) \\ p_{ce}(Y^{t-1}, X^t) &:= q(Y^{t-1}) p(X^t | Y^{t-1}) \\ p_c(Y^{t-1} | X^t) &:= \frac{p_{ce}(Y^{t-1}, X^t)}{\sum_{Y^{t-1} \in \Sigma} p_{ce}(Y^{t-1}, X^t)} \end{aligned} \tag{4}$$

where $q(Y^{t-1})$ is the uniform distribution over the state space of Y. The distribution $p_e(Y^t | X^{t-1})$ indicates the extent to which the current state of Y is an effect caused the previous state of X. Likewise, $p_c(Y^{t-1} | X^t)$ indicates the extent to which the previous state of Y was a cause of the current state of X.

2.2. Brain Process Modules

The brain carries out a wide array of distinct, though integrated processes. While it is difficult to list and classify all of them, they may be roughly partitioned into three general interconnected categories: (i) pre-conscious processes, (ii) conscious processes, and (iii) post-conscious processes.

Pre-conscious processes are those which occur independent of conscious experience. The activity of the autonomic nervous system is paradigmatic of this category. Though extremely important for sustaining life, these functions are somewhat irrelevant to our considerations and shall hence be ignored in what follows.

Conscious processes are those which directly give rise to conscious experience; that is, they govern the dynamics of the neural correlates of consciousness, and include those responsible for perception, the categorical discrimination thereof, awareness, and short-term memory recall, among other things. They are not to be conflated with the first-person subjective conscious *experiences* to which these correlates are thought to somehow give rise. At the physiological level, all we are concerned with are the neural correlates of conscious experience and awareness; we are agnostic as to *how* the mental states are determined by these correlates, and therefore do not commit to any view about the origins of consciousness as such.

Post-conscious processes are those which are not the primary basis for conscious experience, but still depend on the correlates of consciousness such as language processing and inference-making. This class of brain functions is roughly equivalent to cognitive processes[2].

Each of these classes of brain processes has a reasonably well-defined collection of physiological regions in the brain which carry them out. Hence it is possible for us to conceptually partition the brain into three physical "modules." The important feature of these modules is that they are deeply interconnected. While it is hard to cleanly demarcate their physiological boundaries, what is important for our purposes is not how to carve up the brain into these modules, but the causal relations *between* them.

In the SLM (cf. Section 2.1), we shall denote the "consciousness" module by $X_{\text{Con}} \subseteq B$ and the individual connectomic units that compose it by $\{x_i\}$. Likewise, we shall denote the "cognition" module by $Y_{\text{Cog}} \subseteq B$ and the connectomic units that compose it by $\{y_i\}$. The region of the brain that is relevant for our purposes is the joint system $X_{\text{Con}} \cup Y_{\text{Cog}}$.

2.3. Causal Feedback

We now argue that the brain modules X_{Con} and Y_{Cog} mutually exhibit causal feedback.

To see that X_{Con} causally influences Y_{Cog}, note that cognitive tasks are like computational tasks (broadly construed) which take as their inputs the correlates of consciousness. For instance, learning is a cognitive process that is informed by sensory stimuli. Likewise, language processing is a cognitive process that begins with a more abstract input of which the cognizing subject is usually consciously aware. More generally, changing what a person perceives or is conscious of affects how they make sense of their perceptions and what sorts of inferences they will draw.

What does the causal relation from X_{Con} to Y_{Cog} look like? It is known that a single neuron may participate in bringing about many sorts of perceptions and experiences, and many different neuronal states may correspond to one and the same perceptual experience (as there is great degeneracy). Hence, one cannot easily reduce a correlate of consciousness to an arrangement of neurons. That is, the correlates of consciousness are not identical to the state of X_{Con}—they are only determined by X_{Con}. More specifically, the intrinsic network of causal influences within X_{Con} determines these neural correlates (see Tononi and Edelman, 1998; Edelman, 2005; Park and Friston, 2013 for discussion).[3] In order for cognition to take the correlates of consciousness as inputs, the system Y_{Cog} must be connected to system X_{Con} in such a way that the internal causal structure of X_{Con} is "read off" of its state and encoded directly into the states of the neurons of Y_{Cog}, which must encode features of the probability distributions p_e, p_{ce}, and p_c of the subsystem X_{Con}. Since we shall establish that there are causal relations in both directions, to prevent circularity, we suppose that Y_{Cog} represents the intrinsic causal structure of X_{Con} as it appears when marginalized to X_{Con} (i.e., ignoring correlations with Y_{Cog}). Determining exactly how this translation could be carried out would require a full account of the emergence of conscious experience from the relevant causal information which we do not have. However, one may view the

[2] It is worth mentioning that there may be some cognitive processes which may be independent of conscious experience. We consider such processes to be *pre-consicous* ones, and thus irrelevant to our analysis.

[3] While the dynamical evolution of the brain may be reduced to a description of its individual neurons, and while its intrinsic causal structure is grounded in the interactions of these neurons, the intrinsic causal structure is not robust against small changes to the network architecture. It is in this way that neural correlates of consciousness are not "reducible" to individual neurons.

units of Y_{Cog} as "simulating" the intrinsic causal structure of X_{Con}, and then carrying out an effective computing procedure on this simulation—this simulation could be modeled with ideas from hierarchical predictive processing which adopts a similar organizational structuring of the brain (cf. Friston, 2005; Friston and Kiebel, 2009; Clark, 2013). In summary, the causal relation $X_{\text{Con}} \rightarrow Y_{\text{Cog}}$ is highly non-trivial.

What does the causal relation from Y_{Cog} to X_{Con} look like? On its own, the system X_{Con} gives rise to the moment-by-moment passive perceptions present in the thinking subject's conscious experience. However, the content of conscious experience—at least for humans—is not merely a passive stream of perception; there is further underlying semantic content within these perceptions of which we come to be aware by carrying out cognitive tasks. While our perceptual apparatus may be capable of carrying out discrimination tasks to categorize our perceptions (e.g., such that we may become aware of the presence of "pain" or "blue" and so on within a given experience), we also come to be consciously aware of much richer structural and abstract features as well. Deprived of all sensory input, the mathematician may still prove complex theorems structured by a sophisticated underlying mathematical grammar and logic, but only if they are consciously aware that they are doing so. To the extent that the thinking subject may be conscious of the *outcomes* of their cognition—which they certainly are in many cases—we see that there must exist some non-trivial causal relation between Y_{Cog} and X_{Con} in which the former causally influences the latter.

More specifically, acts of cognition may change the content of conscious experience such that we may acquire *understanding* of our perceptions, for instance by giving them grammatical structure (over and above merely discriminating qualia), or by carrying out introspection or higher inference-making. It is through this process that one may go from a state of mind of the form "it is the case that ϕ" to the state of mind "I *know* that it is the case that ϕ." Likewise, it is through this process that one may go from the state of mind that "it is the case that ϕ and $\phi \rightarrow \psi$" to the state of mind "it is the case that ψ" (via inference by *modus ponens*). In short, the outcomes of cognitive processes are *re-integrated* back into the correlates of consciousness. This causal feedback via simulation and re-integration between modules in illustrated in **Figure 2**.

We have established that cognition causally influences the content of conscious experience, and vice versa. This is not to say, however, that cognition is itself "perceived." In everyday life, the content of our experience forms the basis of some cognitive inference we may make and we become aware of the outcome of this inference, but we never perceive the inference itself. Indeed, even when one is proving mathematical theorems, at most one is aware of what cognitive rules they are applying when carrying out a deduction: they do not, however, experience the application of these rules as such. This illustrates that, while we argue that cognition causally influences the course of conscious experience in a very strong way, it is not itself *directly* responsible for conscious experience; the neuronal basis for cognition is not itself populated with correlates of consciousness, it merely interacts with these correlates in a reentrant manner. In this sense, we may faithfully view the cognitive module Y_{Cog} as implementing feed-forward computing procedures, e.g., through a neural network that is reintegrated with X_{Con} (such that inference making *in its entirety* is not merely a computing procedure).

Formally, since both X_{Con} and Y_{Cog} causally influence one another in a highly non-trivial manner, we expect that

$$p(X^{t+1}_{\text{Con}}|Y^{t}_{\text{Cog}}) \neq p(X^{t+1}_{\text{Con}}|X^{t}_{\text{Con}}) \tag{5}$$

$$p(Y^{t+1}_{\text{Cog}}|Y^{t}_{\text{Cog}}) \neq p(Y^{t+1}_{\text{Cog}}|X^{t}_{\text{Con}}). \tag{6}$$

Thus, the simulation of X_{Con} encoded in Y_{Cog} will generally not be a faithful predictor of the future behavior of X_{Con}, since it ignores its own causal influence on this behavior. This is the reason we suppose that Y_{Cog} simulates the causal structure of X_{Con} as marginalized to X_{Con}. In **Box 1** we provide a concrete realization of the SLM presented above model, as well as its application to self-reference.

3. SELF-REFERENCE IN THE STRANGE LOOP MODEL

Here we use the SLM to investigate how to make sense of self-reference by unfolding the inconsistency in time (Section 3.1) and provide some clarifying remarks (Section 3.2). Then we show how logical inconsistency is transformed to incongruence (Section 3.3), and argue that the brain does not get caught in endless loops (Section 3.4).

3.1. Unfolding Self-Reference in Time

We now analyze how the intrinsic thought process of an agent carrying out a self-referential deduction as given by the Inclosure Schema (**Box 2**) would appear in the dynamical behavior of the joint system $X_{\text{Con}} \cup Y_{\text{Cog}}$. In formal logic, a deduction in a given formal system is a sequence of grammatically well-formed strings of symbols such that each string is either an instance of an assumed axiom or premise, or is the result of the application of a permitted rule of inference to previous lines in the sequence. If one views a deduction as a dynamical time-dependent thought process in which each line in the deduction corresponds to some fact about which the thinking subject is aware, the sequential ordering of the lines of the deduction may be interpreted as the time ordering of a series of mental states (and thus, a constraint of the compatible dynamics of the underlying brain states).

Given some statement ϕ, to say that an agent is aware of ϕ at time t is to say that the *physical* state of X^{t}_{Con} grounds the *mental* state of being aware of ϕ. One can actively perceive ϕ by occupying such a mental state, or one can remember having perceived ϕ at a previous time. Thus, there is an internal time index $\tau \leq t$ that tracks the time at which ϕ was perceived that may differ from the time index of the state of X_{Con}. If we denote that class of all brain states that give rise to this mental state by $[\phi]$, and index the time at which ϕ is thought to be (or have been) perceived by $[\phi^{\tau}]$, we thus have $X^{t}_{\text{Con}} \in [\phi^{t}]$ if the thinking subject is actively thinking about ϕ, and $X^{t}_{\text{Con}} \in [\phi^{\tau}]$ for $\tau < t$ if they are recalling having thought about ϕ previously.

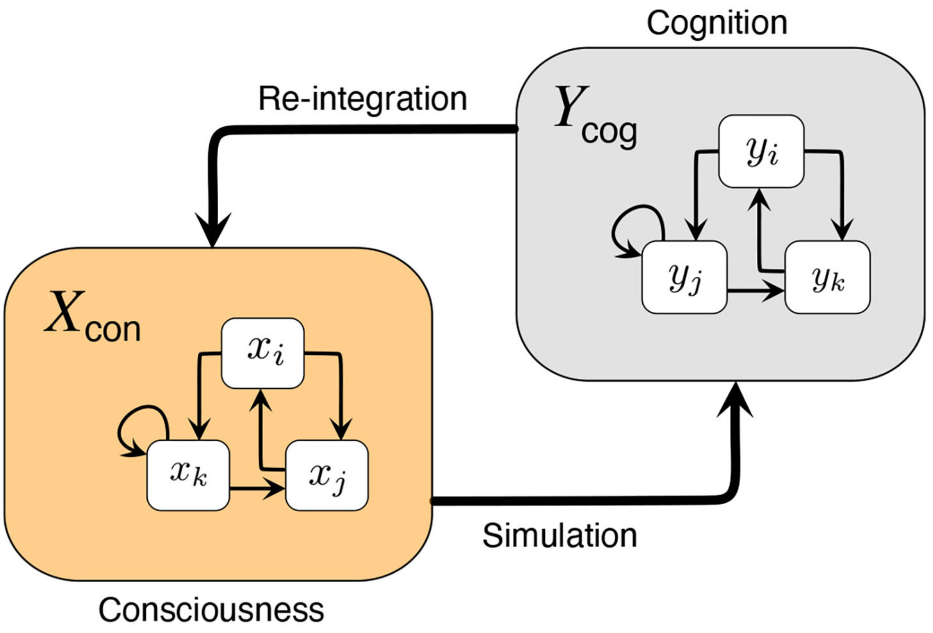


FIGURE 2 | The causal relations $X_{\text{Con}} \to Y_{\text{Cog}}$ needed to simulate perceptions for inference-making, and $Y_{\text{Cog}} \to X_{\text{Con}}$ manifest in the awareness of the outcome of cognitive processing.

BOX 1 | A concrete realization of the Strange Loop Model

To instantiate the SLM, suppose first that a mental state amounts to the awareness of some sentence in a formal language L. Such sentences carry an internal time index τ: at physical time t, one may occupy a mental state of *remembering* some sentence ϕ at an earlier time (i.e., $\tau < t$), they may *anticipate* being aware of ϕ in the future (i.e., $\tau > t$), or they may be aware of ϕ as a feature of the present experience (i.e., $\tau = t$). We suppose that every pair (ϕ, τ) is represented by a unit of Y_{Cog}. The mental state determined by the state of X_{Con} is simulated by the elements of Y_{Cog} via an injective map $S : \Sigma_X \to \{(\phi, \tau) | \phi \in L, \tau \in \mathbb{Z}\}$ where Σ_X is the state space of X_{Con}. That is, S takes the state of X_{Con} to the unit of Y_{Cog} that represents the corresponding mental state. The state of each unit $y = (\phi, \tau) \in Y_{\text{Cog}}$ at time t is given by $y^t = (a^t(y), s^t(y)) \in \{0, 1\} \times \{0, 1\}$, where $a^t(y) = 1$ if the thinking subject is consciously aware of y at time t, and it is 0 otherwise, and $s^t(y) = 1$ if the thinking subject assigns truth to ϕ at time τ (i.e., if they think ϕ was/is/will be true at time τ), and it is 0 otherwise. The state of Y_{Cog} at time t is determined by:

$$y^t = (a^t(y), s^t(y)) = \begin{cases} (1, 1) & \text{if } y = S(X^t_{\text{Con}}) \\ (0, s^{t-1}(y)) & \text{if } y \neq S(X^t_{\text{Con}}). \end{cases}$$

That is, to be aware of (ϕ, τ) at time t is to think it to be true, and to think ϕ is false is to be aware of the truth of $\neg\phi$. The state of Y_{Cog} at time $t + 1$ is determined by the application of some inferential mechanism by Y_{Cog}. If the thinking subject applies a rule of inference of the form $\{\sigma_1, \ldots, \sigma_k\} \vdash \psi$, $a^t(y)$ is updated so that one is only presently aware of ψ, namely $a^{t+1}(\psi, t + 1) = 1$, and $a^{t+1}(\phi, \tau) = 0$ for $\phi \neq \psi$ and any τ. The transition rule for s^t is

$$s^{t+1}(\psi, t + 1) = \prod_{i=1}^{k} s^t(\sigma_i, t)$$

and $s^{t+1}(\phi, \tau) = s^t(\phi, \tau)$ for all τ when ϕ is independent of ψ, and $s^{t+1}(\xi, \tau) = s^t(\xi, \tau)$ for all $\tau \neq t + 1$ and any ξ. Sentences containing ψ have their truth values adjusted according with the change in the truth value of ψ, for example $s^{t+1}(\neg\psi, t + 1) = 1 - s^{t+1}(\psi, t + 1)$ and $s^{t+1}(\phi \wedge \psi, t + 1) = s^{t+1}(\phi, t + 1) \cdot s^{t+1}(\psi, t + 1)$ and so on. We do not fully specify the transition rule for X_{Con}, but we require that it be such that after such an inference, $S(X^{t+1}_{\text{Con}}) = (\psi, t + 1)$.

The self-referential paradox arises when one may assert that $(a^t(\phi, t_1), s^t(\phi, t_1)) = (a^t(\neg\phi, t_2), s^t(\neg\phi, t_2))$ for $t_1 \neq t_2$. But, as shown in Section 3.1, this scenario is not challenging to understand; these are two different nodes of Y^t_{Cog}, and there is no consistency requirement preventing this as a value assignment. Even if one imposes consistency conditions at equal times, since these are unequal-time units, such conditions need not prohibit this behavior.

If ϕ and ψ are two formulas that are not logically equivalent to one another, one might suppose that $[\phi^\tau] \cap [\psi^\tau] = \emptyset$. This very general claim may be objected to in principle by noting, for instance, that if ϕ and ψ are sufficiently complex, the thinking subject may not always be immediately aware of their logical (in)equivalence[4]. Nevertheless, it should be agreeable that there

[4] One might consider, for instance, the classic example due to (Descartes, 1993, Meditation VI) of the chiliagon (a 1000-sided polygon). The thinking subject, it may be argued, uses an identical mental representation to depict a polygon with

BOX 2 | Diagonalization, self-reference and paradoxes

While self-reference and its paradoxical consequences arise in a wide range of settings, the construction of the self-referential statement leading toward contradiction typically has a standard form, termed the Inclosure Schema; cf. Priest, 2002, Chapter 9.4). At a higher level of abstraction, this may be viewed as an instance of Lawvere's Theorem (Yanofsky, 2003; Lawvere, 2006; Roberts, 2021).

In plain language, the relevant actresses of the Inclosure Schema are the following. A predicate is a property that elements of a set may possess, and we identify the predicate with its extension, i.e., with the set of elements that instantiate it. For example, the predicate "odd" of the set of natural numbers is the set $\{1, 3, 5, 7, 9, \ldots\}$. If a set x has property P, we write $P(x)$, meaning that $P(x)$ is true, i.e., x is in the extension of P. We will consider the collection of all sets V, and a function $\Delta : V \to V$.

More formally, let φ and ψ denote two predicates that may apply to arbitrary sets (where "set" is meant in the sense of natural language, which is more expressive than formal set theory at the cost of being inconsistent), and let Δ be a function on sets. Then self-reference occurs when:

1. $E_\varphi = \{y|\varphi(y)\}$ is a set, and $\psi(E_\varphi)$

2. If $x \subseteq E_\varphi$ such that $\psi(x)$, then $\Delta(x) \notin x$ and $\Delta(x) \in E_\varphi$

Statement 1 says that the extension of the predicate φ is a set and is called E_φ, and that E_φ has property ψ. Statement 2 defines the features of Δ, namely Δ takes sets with property ψ whose *elements* all have property φ to sets whose elements have property φ but are not contained in the original set. The contradiction associated with self-reference appears when one applies condition 2 to the maximal subset, namely, E_φ itself, from which it follows that $\Delta(E_\varphi) \in E_\varphi$ and $\Delta(E_\varphi) \notin E_\varphi$; a contradiction.

Let us see this argument in action by considering Russell's paradox. In naïve set theory, the extension of any predicate is a set. Russell's paradox is as follows: suppose X is the set of all sets that do not contain themselves. Then if $X \in X$, by definition it follows that $X \notin X$. However, if $X \notin X$, then since X is the set of all sets that do not contain themselves, we find $X \in X$; a contradiction. On the Inclosure Schema this paradox may be recast as follows. First, ψ is the predicate "is a set," φ is the predicate "does not contain itself," and Δ is defined by $\Delta(x) = \{y \in x | y \notin y\}$, i.e., it's image is the set of all sets in x that do not contain themselves. Since ψ is a predicate in naïve set theory, 1 is true and asserts that E_φ exists and is the notorious "set of all sets that do not contain themselves." Then if x is a set, clearly $\Delta(x) \in E_\varphi$. Likewise, if $\Delta(x) \in \Delta(x)$, then by definition of Δ, $\Delta(x) \notin \Delta(x)$ and so we must conclude $\Delta(x) \notin x$. Thus 2 is also satisfied. But then setting $x = E_\varphi$, this implies simultaneously that $\Delta(E_\varphi) \in E_\varphi$ and $\Delta(E_\varphi) \notin E_\varphi$; a contradiction. This contradiction historically called for the reformulation set theory and was one of the many factors leading to modern-day ZF axiomatic set theory. All other famous self-reference paradoxes may be articulated using this Inclosure Schema.

are no brain states that are simultaneously neural correlates of the awareness of ϕ and also neural correlates of the awareness of $\neg\phi$. This weaker hypothesis is all we shall require. Then, if the thinking agent carries out a deductive inference whose sequential lines are denoted $\{\phi_n\}$, this corresponds to their brain undergoing a dynamical evolution of the form:

$$X_{\text{Con}}^{t} \in [\phi_0^{\tau(t)}] \to X_{\text{Con}}^{t+1} \in [\phi_1^{\tau(t+1)}] \to X_{\text{Con}}^{t+2} \in [\phi_2^{\tau(t+2)}] \to \cdots \to X_{\text{Con}}^{t+n} \in [\phi_n^{\tau(t+n)}] \quad (7)$$

where $\tau : \mathbb{Z} \to \mathbb{Z}$ satisfies $\tau(t) \leq t$. While the individual lines of a deduction correspond to mental states (and thus restricted classes of brain states), the axioms and rules of inference from which subsequent lines are produced do not reflect processes of which one is consciously aware during such a thought process. Rather, they reflect the cognitive rules that the thinking agent's brain may apply to the content of their experience in order to bring about their subsequent mental states. In this way, the axioms and rules of inference that enable one to formalize a given deduction correspond in the underlying thought process to processes implementations of cognitive processes via Y_{Cog} (see **Box 1** for a concrete realization thereof).

To illustrate this, let us consider a simple example. Suppose one sees a green apple before them. This perception, and the discrimination of various features of this perception are grounded in neural correlates that reside physiologically in the brain module X_{Con} at the present time t. Suppose, subsequently (say, at time $t + 1$), that one remembers from their past experiences that essentially all green apples have a sour taste. (Of course, the inductive formation of such a generalized belief from past memories is non-trivial, but it nevertheless happens.) This association, then, of sour flavor with green apples in general is something about which the thinking subject becomes consciously aware, and hence forms part of their conscious experience. Therefore, it is likewise encoded in the neural correlates of consciousness present in X_{Con} at time $t + 1$. From these two perceptions, the thinking subject may apply *modus ponens* to conclude that the apple they saw at time t would likely have had a sour taste were they to eat it. The general rule of *modus ponens*, however, is not something of which one has direct perception when it is being implemented; making such inferences is a higher cognitive process. The implementation of *modus ponens*, therefore, is a process carried out by the brain module Y_{Cog}. Importantly, once this inference has been carried out, the subject becomes aware of its outcome. Namely, at a subsequent time (say, $t + 2$), they become consciously aware that, had they eaten the apple, it would likely have tasted sour. This is the general manner in which deduction may be realized as thought processes implemented within our brain model.

We now apply this perspective to the linguistic processing of self-referential statements via the Inclosure Schema (see **Box 2**). The idea is to distinguish between the abstract logical results and the thought processes obtained when a thinking subject confronts an instance of self-reference and thinks about it over a finite period of time. Logically speaking, the contradiction arising

1000 sides as one with only 999 sides. Such mental representations, then, would be given the same class of corresponding brain states, even if it is a logically different object. Hence, logical inequivalence is inadequate for individuating mental states (or brain states).

from a diagonalization argument is absolute; we do not contest this. However, when we infer this contradiction—i.e., when the dynamical behavior of a subject's brain implements the thought process that yields this contradiction—using diagonalization, we do so in two temporally separate parts; first, we prove that $\Delta(x) \notin x$ and conclude that $\Delta(E_\varphi) \notin E_\varphi$. Then, at a later time, we conclude that $\Delta(E_\varphi) \in E_\varphi$. The contradiction arises when we remember at a third time that we had proven both of these two facts separately.

Let us look at Tarski's paradox to see this play out concretely, following the exposition by Priest (2002). To begin, let T be a "truth" predicate on sentences, i.e., for any sentence x, $T(x)$ is true if and only if x is true (this is called Tarski's *T-schema*). Let ψ denote definability such that $\psi(X)$ is true for any set of sentences X just in case there exists a sentence x which defines X as a set (of sentences). If X is any definable set of sentences, let $\Delta(X) = \alpha$ where $\alpha = \langle \alpha \notin X \rangle$ (here $\langle \cdot \rangle$ is used to denote the proper name of a sentence). That is, $\Delta(X)$ is the sentence α which expresses that α is not an element of the set of sentences X. Clearly, α is self-referential. If an agent thinks about the T-schema, their thought process might look like the following. First, one supposes that the totality of all true sentences exists and is definable, that is, that $\mathrm{Tr} := \{x \mid T(x)\}$ is a set that may be defined by some sentence. If X is definable (whence $\psi(X)$ is true) and if $X \subseteq \mathrm{Tr}$, we have in the temporal framework described:

Time	Inference	Rule
$t = 0$	$\Delta(X) \in X \rightarrow \langle \alpha \notin X \rangle \in X$	Definition of Δ
$t = 1$	$\langle \alpha \notin X \rangle \in X \rightarrow \langle \alpha \notin X \rangle \in \mathrm{Tr}$	Comprehension in ZF
$t = 2$	$\langle \alpha \notin X \rangle \in \mathrm{Tr} \rightarrow \alpha \notin X$	T-Schema
$t = 3$	$\alpha \notin X \rightarrow \Delta(X) \notin X$	Definition of Δ
$t = 4$	$\Delta(X) \in X \rightarrow \Delta(X) \notin X$	Modus ponens (three times)
$t = 5$	$\Delta(X) \notin X \rightarrow \Delta(X) \notin X$	Tautology
$t = 6$	$(\Delta(X) \in X) \vee (\Delta(X) \notin X) \rightarrow \Delta(X) \notin X$	Propositional logic
$t = 7$	$(\Delta(X) \in X) \vee (\Delta(X) \notin X)$	Excluded middle
$t = 8$	$\Delta(X) \notin X$	Modus ponens on $t = 6$ and $t = 7$
$t = 9$	$\Delta(\mathrm{Tr}) \notin \mathrm{Tr}$	Substitution of $X = \mathrm{Tr}$ to $t = 8$
$t = 10$	$\Delta(\mathrm{Tr}) \in \mathrm{Tr}$	Substitution of $X = \mathrm{Tr}$ to $t = 1$
$t = 11$	$(\Delta(\mathrm{Tr}) \in \mathrm{Tr}) \wedge (\Delta(\mathrm{Tr}) \notin \mathrm{Tr})$	Propositional logic

Let us now look at the brain states that could in principle produce the mental states associated with each line of this deduction. We may rewrite the above inference as follows:

$$X^0_{\mathrm{Con}} \in [(\Delta(X) \in X \rightarrow \langle \alpha \notin X \rangle \in X)^0]$$
$$X^1_{\mathrm{Con}} \in [(\langle \alpha \notin X \rangle \in X \rightarrow \langle \alpha \notin X \rangle \in \mathrm{Tr})^1]$$
$$X^2_{\mathrm{Con}} \in [(\langle \alpha \notin X \rangle \in \mathrm{Tr} \rightarrow \alpha \notin X)^2]$$
$$X^3_{\mathrm{Con}} \in [(\alpha \notin X \rightarrow \Delta(X) \notin X)^3]$$
$$X^4_{\mathrm{Con}} \in [(\Delta(X) \in X \rightarrow \Delta(X) \notin X)^4]$$
$$X^5_{\mathrm{Con}} \in [(\Delta(X) \notin X \rightarrow \Delta(X) \notin X)^5]$$
$$X^6_{\mathrm{Con}} \in [((\Delta(X) \in X) \vee (\Delta(X) \notin X) \rightarrow \Delta(X) \notin X)^6]$$
$$X^7_{\mathrm{Con}} \in [((\Delta(X) \in X) \vee (\Delta(X) \notin X))^7]$$

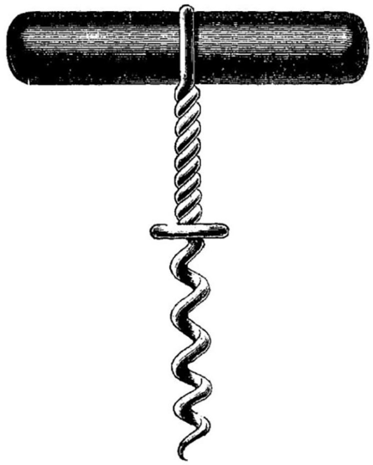


FIGURE 3 | Unfolding self-reference in time can be imagined as unfolding a circle many-times packed into a corkscrew, where the time dimension corresponds to the long dimension of the corkscrew. Equivalently, it can be imagined as the evolution of circularly polarised light.

$$X^8_{\mathrm{Con}} \in [(\Delta(X) \notin X)^8]$$
$$X^9_{\mathrm{Con}} \in [(\Delta(\mathrm{Tr}) \notin \mathrm{Tr})^9]$$
$$X^{10}_{\mathrm{Con}} \in [(\Delta(\mathrm{Tr}) \in \mathrm{Tr})^{10}]$$
$$X^{11}_{\mathrm{Con}} \in [((\Delta(\mathrm{Tr}) \in \mathrm{Tr})^9 \wedge (\Delta(\mathrm{Tr}) \notin \mathrm{Tr})^{10})^{11}]$$

To prove a contradiction in time in a manner that could require a physically impossible brain state, one would need to show that $X^t_{\mathrm{Con}} \in [\phi^t]$ and $X^t_{\mathrm{Con}} \in [\neg\phi^t]$ for a single t. This does not happen. In this way, if we want to model deductive inferences as processes carried out by a physical systems such as the brain which evolves in time, we see that the contradictions appear not directly, but spread out in time and then recalled, and so they may be implemented by a machine such as the brain that operates in time (**Figure 3**). In particular, we do not encounter the fractal picture given in Grim et al. (1993).

Moreover, because it is possible to have $X^t_{\mathrm{Con}} \in [\phi^t]$ and $X^{t'}_{\mathrm{Con}} \in [\neg\phi^{t'}]$ at different times $t \neq t'$, we see that the brain has on this model sufficient expressive power to treat truth values as dynamically changing quantities. This may be contrasted with Turing machines tasked with deciding truth values; the state of such a machine may evolve in time, but the truth value it aims to decide is static.

3.2. Clarifying Remarks

Let us make a few remarks on the conclusions reached so far. We are not denying the logical contradiction that appears in the above deduction. Indeed, what we have done here amounts to a temporal version of what (Priest, 2002) calls parameterization; it is a standard approach to avoid paradoxes, and in general, any contradiction that is avoided by parameterization will reappear at a higher level again when one analyzes the parameterized formalism. However, this is irrelevant to our aims: what we have shown is that an inference-making device that has a register that expresses its state of deduction in time (while some auxiliary system carries out further inference-making tasks leading to eventual update of the register) can effectively model contradictory scenarios without existing in a contradictory state itself. That is, there is never an instant where such a system need occupy two different physical states simultaneously.

Extending this to our model of the brain, the "inference" column label could be replaced with "the thought of which the conscious agent is aware" at each given time, while the "rule" column label could just as well be interpreted as "the cognitive process being carried out in the intermediate time window." In this way, we have a rough picture for how the brain could physically model the contradictions that arise from self-reference paradoxes (noting that the above proof for the contradiction in Tarski's paradox is of the generic diagonalization form) without itself being in any strange superposition of disagreeing physical configurations.

What makes this temporal parameterization technique useful is that while in a purely logical setting, the relation between subsequent lines in a deduction is strictly a logical one (with no temporality and so forth), when represented on a physical system, is no longer an abstractly logic relation, but is instead a *causal* relation indicating an interaction between these two brain modules we have discussed. In particular, it is a causal relation which requires an intermediate physical process to commence and terminate. Hence, there is an intermittent time, and so the contradiction may be "stretched out" in time in the appropriate sense. (This is analogous to the Kantian view of time as a means for the thinking subject to experience contradictory perceptions without an actual contradiction obtaining Kant, 1998, A32/B48).

3.3. Transforming Logical Inconsistency to Incongruence

We now apply the IIT formalism (see **Box 3**) to the SLM, and show how the logical inconsistency of self-referential paradoxes is transformed to incongruence.

First observe that since the correlates of consciousness were taken to reside in X_{Con}, it is reasonable to suppose that for any subsystem of the brain $Z \subseteq B$, if Z is maximally irreducible while in some state Z^t, it must be the case that $Z \cap X_{\text{Con}} \neq \emptyset$. In most cases, Z will simply be a subsystem of X_{Con}. However, from Equation (5), there will be some irreducible subsystems that overlap with Y_{Cog} as well. In particular, $X_{\text{Con}} \cup Y_{\text{Cog}}$ is expected to be maximally irreducible.

Incongruence in IIT is defined as follows. For any system S, given a pair of subsystems $G, H \subseteq S$, G and H are incongruent if they make differing predictions about the past or future behavior of some particular node $z \in S$ (see Haun and Tononi, 2019; Albantakis and Tononi, 2019, p. 5). This occurs, for instance, if $p(z^{t+1}|G^t) \neq p(z^{t+1}|H^t)$. When self-referential inferences are made, if we suppose ϕ is thought about at time t, $\neg\phi$ is thought about at $t+1$, and $\phi^t \wedge \neg\phi^{t+1}$ is thought about at time $t+2$, then it is because of the cognitive processes in Y_{Cog} implemented at time t and $t+1$ that this is the case. In particular, if we presently think some sentence is true, we expect that it will be true still at the next instant, so that

$$p(X_{\text{Con}}^{t+1} \in [\neg\phi^{t+1}] \mid X_{\text{Con}}^{t} \in [\phi^t], Y_{\text{Cog}})$$

is large, while

$$p(X_{\text{Con}}^{t+1} \in [\neg\phi^{t+1}] \mid X_{\text{Con}}^{t} \in [\phi^t])$$

is small. However, Y_{Cog} implements a rule of inference in this transition, which causes $X_{\text{Con}}^{t+1} \in [\neg\phi^{t+1}]$ to occur. During self-referential inferences, not only do two different subsystems disagree about the probabilities assigned to a particular node's future state (cf. Equation 5); rather, they assign essentially *opposite* probabilities to the future behavior of the subsystem spanned by all maximally irreducible subsystems. Hence, incongruence arises in a strong way.

Put differently, causal incongruence in IIT offers a precise sense in which the parts of a system fail to describe the whole of the system, namely, taken separately, the parts may disagree with one another about the descriptions they provide. In the SLM framework, this is exploited as a *feature*: it is this disagreement that enables the brain to represent contradictions in the requisite manner needed to make sense of self-referential statements.

3.4. Avoiding Unhalting Cycles

We now argue that the cyclic behavior of the SLM, as described in Section 3.1, does not persist *indefinitely* (as it would for an unhalting Turing machine). When the thinking subject gets caught in a cognitive cycle of the same form, if their attention is drawn away from the cyclic inference at hand, the cycle will end. This is so because, as a thinking subject learns by repeating a task many times, they devote less and less attention and focus toward the task being learned (Kandel et al., 2013, Chapter 64). In the present context, this means that if the thinking subject cycles through the thought process associated with deriving disagreeing truth values for a self-referential statement, they will not get caught in a loop, but rather will pay less attention to the inference upon subsequent iterations. Since the brain actively monitors a large class of sensory stimuli and implements many cognitive processes in parallel, as this attention diminishes, the thinking subject is increasingly likely to refocus their attention elsewhere. In short, if attention is a resource, the architecture of the brain is such that the re-allocation of this resource inhibits the ensuing feedback and makes infinite inferential loops unstable.

This is analogous to binocular rivalry, where the subject's visual field is eventually changed, whence their visual sensations escape from flowing toward lock-in states (Hohwy et al., 2008; Clark, 2013), and to visual paradoxes, like the Necker cube (where two alternative possible attractors are present) or the recognition of ambiguous images (Inoue and Nakamoto, 1994; Kelso, 1995). This metastable behavior due to self-reference can also be found in gene networks, where the causal feedback associated with cross-regulatory interactions

BOX 3 | Integrated Information Theory

Integrated Information Theory (IIT; see Oizumi et al., 2014; Tononi and Koch, 2015; Tononi et al., 2016) is a framework that seeks to provide a constructive account of the origins of conscious experience by describing it as an emergent feature of causally integrated dynamical systems such as the brain. IIT begins by articulating those features of conscious experience that one might take to be constitutive, and then identifies features of the causal structure of a dynamical system that qualitatively realize these features (in a manner that can be made quantifiably precise via informational measures). A model of the IIT formalism is a dynamical system X (such as the SLM) together with all of the probabilities of the form $p(x_i^{t+1}|X^t)$. From the causal probabilities defined in Equation (4), the IIT formalism defines measures to quantify the extent to which a subsystem $S \subseteq X$ cannot be causally reduced, e.g., to a pair of subsystems G and H with $S = G \cup H$ and the extent to which every such S is causally integrated.

A state of a subsystem at some time is *irreducible* if and only if the probabilities that characterize its intrinsic causal structure cannot be exactly recovered by partitioning it into subsystems. Irreducibility is quantified using an informational measure; those subsystems of X that realize the maximum of this measure for the system X are called *maximally irreducible*. There are generally many different maximally irreducible subsystems.

According to IIT, only those subsystems that are maximally irreducible at a given moment contribute to consciousness at that moment, forming the instantaneous correlates of consciousness. The manner in which a maximally irreducible subsystem contributes to consciousness is dictated by its causal probabilities which populate points in a supposed space of qualia. The conscious experience realized by a physical substrate (a human brain or otherwise) is a byproduct of that substrate's maximally irreducible intrinsic causal structure. Here we do not assess the plausibility of IIT as a theory of consciousness; rather, we note that our SLM can be recast within the IIT formalism straightforwardly.

can be spread in time or space leading to interesting phenomena (Isalan, 2009).

4. CONCLUSIONS AND OUTLOOK

In this work, we have constructed a high-level discrete dynamical model of the brain, termed the Strange Loop Model (SLM; Section 2), in order to describe inference-making, which uses causal feedback between conscious and cognitive processes. We have used the SLM to model self-reference and shown that logical inconsistencies unfold in time (Section 3.1), and hence the contradictions dissolve, as one never encounters inconsistent truth values simultaneously. Rather, one deduces at different times that a sentence has different truth values and then remembers having carried out both such deductions. This flexibility enables the human brain to model self-reference in a manner that is inaccessible to usual computing devices by construction. We have also applied the SLM within the context of IIT and shown that logical inconsistencies are transformed into incongruences (Section 3.3). Finally we have argued that, because the brain is receptive to a wide range of different stimuli, and because one devotes less attention to repetitive cognitive tasks as time passes, these cyclic inferences are unstable are thus terminated (Section 3.4).

The interaction between X_{Con} and Y_{Cog} via the described causal feedback enables the human mind to be aware of the outcomes of cognitive inferences, and likewise further cognize about such an awareness. Put differently, the causal feedback here described enables the thinking subject to be aware of their own cognitive processes, and to then make inferences about their own cognition. This situation is reminiscent of universality encountered in Turing machines, spin models and neural networks (De las Cuevas, 2020).

Finally, we may compare the SLM with a Turing machine or any other standard computing machine. Unlike an algorithm running in a Turing machine, the processing carried out by the SLM is not a deciding process, because it need not reach a static truth value of a variable. Moreover, the only relevant features of a Turing machine are its input–output functionality (that is, the formal language it recognizes Kozen, 1997), whereas the intrinsic causal structure of the brain is crucial. In this way, we conclude that the process carried out by the brain and the computer is different.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

PF is supported in part by funding from the Social Sciences and Humanities Research Council. RS thanks the Spanish Ministry of Economy and Competitiveness, grant PID2019-111680GB-I00, an AGAUR FI 2018 grant, and the Santa Fe Institute. GD acknowledges funding from the Austrian Science Fund (FWF) (START Prize Y-1261-N).

ACKNOWLEDGMENTS

We wish to thank Larissa Albantakis and Wolfram Hinzen for insightful discussions.

REFERENCES

Albantakis, L., and Tononi, G. (2019). Causal composition: structural differences among dynamically equivalent systems. *Entropy* 21, 989. doi: 10.3390/e21100989

Allman, J. M. (1999). *Evolving Brains*. New York, NY: Scientific American Library.

Arbib, M. (2012). *Brains, Machines, and Mathematics*. New York, NY: Springer.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci*. 36, 181–204. doi: 10.1017/S0140525X12000477

Cobb, M. (2020). *The Idea of the Brain: The Past and Future of Neuroscience*. Hachette: Basic Books.

De las Cuevas, G. (2020). *Universality everywhere Implies Undecidability* Everywhere. FQXi essay.

Descartes, R. (1993). *Meditations on First Philosophy*. Indianapolis, IN: Hackett Publishing.

Edelman, G. (1992). *Bright Air, Brilliant Fire: On the Matter of the Mind*. New York, NY: Basic Books.

Edelman, G. M. (2005). *Wider Than the Sky: The Phenomenal Gift of Consciousness*. New Haven, CT: Yale University Press.

Friston, K. (2005). A theory of cortical responses. *Philos. Trans. R. Soc. B* 360, 815–836. doi: 10.1098/rstb.2005.1622

Friston, K., and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. B* 364, 1211–1221. doi: 10.1098/rstb.2008.0300

Fukushima, K. (1988). Neocognitron: a hierarchical neural network capable of visual pattern recognition. *Neural Netw*. 1, 119–130. doi: 10.1016/0893-6080(88)90014-7

Grim, P. (1993). "Self-reference, chaos, and fuzzy logic," in *Integration of Fuzzy Logic and Chaos Theory*. (Berlin; Heidelberg; New York, NY: Springer), 317–359.

Grim, P., Mar, G., Neiger, M., and St. Denis, P. (1993). Self-reference and paradox in two and three dimensions. *Comput. Graphics* 17, 609–612. doi: 10.1016/0097-8493(93)90013-Y

Haun, A., and Tononi, G. (2019). Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy* 21, 1160. doi: 10.3390/e21121160

Hofstadter, D. (1979). *Gödel, Escher, Bach*. New York, NY: Basic Books.

Hofstadter, D. (2007). *I am a Strange Loop*. New York, NY: Basic Books.

Hohwy, J., Roepstorff, A., and Friston, K. (2008). Predictive coding explains binocular rivalry: an epistemological review. *Cognition* 108, 687–701. doi: 10.1016/j.cognition.2008.05.010

Indiveri, G., and Liu, S. (2015). Memory and information processing in neuromorphic systems. *Proc. IEEE* 103:1379–1397. doi: 10.1109/JPROC.2015.2444094

Inoue, M., and Nakamoto, K. (1994). Dynamics of cognitive interpretations of a necker cube in a chaos neural network. *Progr. Theor. Phys*. 92, 501–508. doi: 10.1143/PTP.92.501

Isalan, M. (2009). Gene networks and liar paradoxes. *Bioessays* 31, 1110–1115. doi: 10.1002/bies.200900072

Jonas, E., and Kording, K. (2017). Could a neuroscientist understand a microprocessor? *PLoS Comput. Biol*. 13:e1005268. doi: 10.1371/journal.pcbi.1005268

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., and Hudspeth, A. J. (Eds.). (2013). *Principles of Neural Science, 5th Edn*. New York, NY: McGraw Hill Medical.

Kant, I. (1998). *Critique of Pure Reason. The Cambridge Edition of the Works of Immanuel Kant*. Cambridge: Cambridge University Press.

Kelso, J. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. Cambridge, MA: MIT Press.

Kozen, D. C. (1997). *Automata and Computability*. New York, NY: Springer.

Krohn, S., and Ostwald, D. (2017). Computing integrated information. *Neurosci. Consciousness* 2017:nix017. doi: 10.1093/nc/nix017

Lawvere, F. W. (2006). "Diagonal arguments and cartesian closed categories," in *Reprints in Theory and Applications of Categories*, Reprints Theory Appl. Categ. 1–13.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Markovi, D., Mizrahi, A., Querlioz, D., and Grollier, J. (2020). Physics for neuromorphic computing. *Nat. Rev. Phys*. 2, 499–510. doi: 10.1038/s42254-020-0208-2

Martinez, P., and Sprecher, S. (2020). Of circuits and brains: the origin and diversification of neural architectures. *Front. Ecol. Evolut*. 8:82. doi: 10.3389/fevo.2020.00082

McCulloch, W., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys*. 5, 115–133. doi: 10.1007/BF02478259

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput. Biol*. 10:e1003588. doi: 10.1371/journal.pcbi.1003588

Park, H.-J., and Friston, K. (2013). Structural and functional brain networks: from connections to cognition. *Science* 342:1238411. doi: 10.1126/science.1238411

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference, 2nd Edn*. Cambridge: Cambridge University Press.

Priest, G. (2002). *Beyond the Limits of Thought*. Oxford: Clarendon Press.

Priest, G. (2006). *In Contradiction*. Oxford: Clarendon Press.

Prokopenko, M., Harre, M., Lizier, J., Boschetti, F., Peppas, P., and Kauffman, S. (2019). Self-referential basis of undecidable dynamics: from the liar paradox and the halting problem to the edge of chaos. *Phys. Life Rev*. 31, 134–156. doi: 10.1016/j.plrev.2018.12.003

Rashevsky, N. (1960). *Mathematical Biophysics: Physico-Mathematical Foundations of Biology, 3rd Edn*. New York City, NY: Dover P. Inc.,.

Roberts, D. M. (2021). Substructural fixed-point theorems and the diagonal argument: theme and variations. *arXiv:2110.00239*.

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Smee, A. (1850). *Instinct and Reason: Deduced From Electro-Biology*. London: Reeve and Benham.

Sporns, O., Tononi, G., and Ktter, R. (2005). The human connectome: a structural description of the human brain. *PLoS Comput. Biol*. 1:e42. doi: 10.1371/journal.pcbi.0010042

Tegmark, M. (2000). Importance of quantum decoherence in brain processes. *Phys. Rev. E* 61, 4194–4206. doi: 10.1103/PhysRevE.61.4194

Tognoli, E., and Kelso, J. A. S. (2014). The metastable brain. *Neuron* 81, 35–48. doi: 10.1016/j.neuron.2013.12.022

Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Rev. Neurosci*. 17, 450–461. doi: 10.1038/nrn.2016.44

Tononi, G., and Edelman, G. M. (1998). Consciousness and complexity. *Science* 282, 1846–1851. doi: 10.1126/science.282.5395.1846

Tononi, G., and Koch, C. (2015). Consciousness: here, there and everywhere? *Philos. Trans. R. Soc. B* 370:20140167. doi: 10.1098/rstb.2014.0167

Wood, G. (2002). *Living Dolls: A Magical History of the Quest for Mechanical Life*. London: Faber & Faber.

Yanofsky, N. S. (2003). A universal approach to self-referential paradoxes, incompleteness and fixed points. *Bull. Symbolic Logic* 9, 362–386. doi: 10.2178/bsl/1058448677

Zylberberg, A., Dehaene, S., Roelfsema, P. R., and Sigman, M. (2011). The human Turing machine: a neural framework for mental programs. *Trends Cogn. Sci*. 15, 293–300. doi: 10.1016/j.tics.2011.05.007

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Ecology and Evolution

ORIGINAL RESEARCH
published: 28 January 2022
doi: 10.3389/fevo.2021.806283

How Organisms Come to Know the World: Fundamental Limits on Artificial General Intelligence

Andrea Roli[1,2]*, *Johannes Jaeger*[3]* and *Stuart A. Kauffman*[4]

[1] Department of Computer Science and Engineering, Campus of Cesena, Università di Bologna, Bologna, Italy, [2] European Centre for Living Technology, Venezia, Italy, [3] Complexity Science Hub (CSH) Vienna, Vienna, Austria, [4] Institute for Systems Biology, Seattle, WA, United States

Artificial intelligence has made tremendous advances since its inception about seventy years ago. Self-driving cars, programs beating experts at complex games, and smart robots capable of assisting people that need care are just some among the successful examples of machine intelligence. This kind of progress might entice us to envision a society populated by autonomous robots capable of performing the same tasks humans do in the near future. This prospect seems limited only by the power and complexity of current computational devices, which is improving fast. However, there are several significant obstacles on this path. General intelligence involves situational reasoning, taking perspectives, choosing goals, and an ability to deal with ambiguous information. We observe that all of these characteristics are connected to the ability of identifying and exploiting new affordances—opportunities (or impediments) on the path of an agent to achieve its goals. A general example of an affordance is the use of an object in the hands of an agent. We show that it is impossible to predefine a list of such uses. Therefore, they cannot be treated algorithmically. This means that "AI agents" and organisms differ in their ability to leverage new affordances. Only organisms can do this. This implies that true AGI is not achievable in the current algorithmic frame of AI research. It also has important consequences for the theory of evolution. We argue that organismic agency is strictly required for truly open-ended evolution through radical emergence. We discuss the diverse ramifications of this argument, not only in AI research and evolution, but also for the philosophy of science.

Keywords: artificial intelligence (AI), universal turing machine, organizational closure, agency, affordance, evolution, radical emergence, artificial life (ALife)

OPEN ACCESS

Edited by:
Giorgio Matassi,
FRE3498 Ecologie et dynamique des systámes anthropiss (EDYSAN),
France

Reviewed by:
Terrence William Deacon,
University of California, Berkeley,
United States
Kalevi Kull,
University of Tartu, Estonia

****Correspondence:***
Andrea Roli
andrea.roli@unibo.it
Johannes Jaeger
jaeger@csh.ac.at

Specialty section:
This article was submitted to Models in Ecology and Evolution, a section of the journal Frontiers in Ecology and Evolution

Received: *31 October 2021*
Accepted: *30 December 2021*
Published: *28 January 2022*

Citation:
Roli A, Jaeger J and Kauffman SA (2022) How Organisms Come to Know the World: Fundamental Limits on Artificial General Intelligence. Front. Ecol. Evol. 9:806283. doi: 10.3389/fevo.2021.806283

1. INTRODUCTION

Since the founding Dartmouth Summer Research Project in 1956 (McCarthy et al., 1955), the field of artificial intelligence (AI) has attained many impressive achievements. The potential of automated reasoning, problem solving, and machine learning has been unleashed through a wealth of different algorithms, methods, and tools (Russell and Norvig, 2021). Not only do AI systems accomplish to perform intricate activities, *e.g.,* playing games (Silver et al., 2016), and to plan complex tasks (LaValle, 2006), but most current apps and technological devices are equipped with some AI component. The impressive recent achievements of machine learning (Domingos, 2015) have greatly extended the domains in which AI can be applied, from machine translation to automatic speech recognition. AI is becoming ubiquitous in our lives. In addition, AI methods are able to produce some kinds of creative artworks, such as paintings (Hong and Curran, 2019), and music (Briot and Pachet, 2020); moreover, GPT-3, the latest version of a deep learning system able to generate texts characterized by surprising writing abilities, has recently been released (Brown et al., 2020) surrounded by some clamor (Chalmers, 2020; Marcus and Davis, 2020).

These are undoubtedly outstanding accomplishments. However, each individual success remains limited to quite narrowly defined domains. Most of today's AI systems are target-specific: an AI program capable of automatically planning tasks, for example, is not usually capable of recognizing faces in photographs. Such specialization is, in fact, one of the main elements contributing to the success of these systems. However, the foundational dream of AI—featured in a large variety of fantastic works in science-fiction—is to create a system, maybe a robot, that incorporates a wide range of adaptive abilities and skills. Hence, the quest for *Artificial General Intelligence (AGI)*, computational systems able to connect, integrate, and coordinate these various capabilities. In fact, true *general intelligence* can be defined as the ability of combining "analytic, creative, and practical intelligence" (Roitblat, 2020, page 278). It is acknowledged to be a distinguishing property of "natural intelligence," for example, the kind of intelligence that governs some of the behavior of humans as well as other mammalian and bird species.

If one considers the human brain as a computer—and by this we mean some sort of computational device equivalent to a universal Turing machine—then the achievement of AGI might simply rely on reaching a sufficient level of intricacy through the combination of different task-solving capabilities in AI systems. This seems eminently feasible—a mere extrapolation of current approaches in the context of rapidly increasing computing power—even though it requires not only the combinatorial complexification of the AI algorithms themselves, but also of the methods used to train them. In fact, many commentators predict that AGI is just around the corner, often admonishing us about the great (even existential) potentials and risks associated with this technological development (see, for example, Vinge, 1993; Kurzweil, 2005; Yudkowsky, 2008; Eden et al., 2013; Bostrom, 2014; Shanahan, 2015; Chalmers, 2016; Müller and Bostrom, 2016; Ord, 2020).

However, a number of serious problems arise when considering the higher-level integration of task-solving capabilities. All of these problems are massively confounded by the fact that real-world situations often involve information that is irrelevant, incomplete, ambiguous, and/or contradictory. First, there is the formal problem of choosing an appropriate metric for success (a cost or evaluation function) according to context and the task at hand. Second, there is the problem of identifying worthwhile tasks and relevant contextual features from an abundance of (mostly irrelevant) alternatives. Finally, there is the problem of defining what is worthwhile in the first place. Obviously, a truly general AI would have to be able to identify and refine its goals autonomously, without human intervention. In a quite literal sense, it would have to know what it wants, which presupposes that it must be capable of wanting something in the first place.

The problem of machine wanting has often been linked by philosophers to arguments about cognition, the existence of subjective mental states and, ultimately, to questions about consciousness. A well-known example is John Searle's work on minds and AI (see, for example, Searle, 1980, 1992). Other philosophers have attempted to reduce machine wanting to cybernetic goal-seeking feedback (e. g., McShea, 2012, 2013, 2016). Here, we take the middle ground and argue that the problem is rooted in the concept of organismic agency, or *bio-agency* (Moreno and Etxeberria, 2005; Barandiaran et al., 2009; Skewes and Hooker, 2009; Arnellos et al., 2010; Campbell, 2010; Arnellos and Moreno, 2015; Moreno and Mossio, 2015; Meincke, 2018). We show that the term "agency" refers to radically different notions in organismic biology and AI research.

The organism's ability to act is grounded in its functional organization, which grants it a certain autonomy (a "freedom from immediacy") (Gold and Shadlen, 2007). An organism not only passively reacts to environmental inputs. It can initiate actions according to internal goals, which it seeks to attain by leveraging opportunities and avoiding obstacles it encounters in its *umwelt*, that is, the world as perceived by this particular organism (Uexküll von, 2010; Walsh, 2015). These opportunities and obstacles are *affordances*, relations between the living agential system and its umwelt that are relevant to the attainment of its goals (Gibson, 1966). Organismic agency enables a constructive dialectic between an organism's goals, its repertoire of actions, and its affordances, which all presuppose and generate each other in a process of constant emergent co-evolution (Walsh, 2015).

Our argument starts from the simple observation that the defining properties of natural systems with general intelligence (such as organisms) require them to take advantage of affordances under constraints given by their particular motivations, abilities, resources, and environments. In more colloquial terms, general intelligences need to be able to invent, to improvise, to *jury-rig* problems that are relevant to their goals. However, AI agents (unlike biological ones) are defined as sophisticated *algorithms* that process information from percepts (inputs) obtained through sensors to actions (outputs) implemented by effectors (Russell and Norvig, 2021). We elaborate on the relation between affordances and algorithms—defined as computational processes that can run on universal Turing machines—ultimately arriving at the conclusion that identifying and leveraging affordances goes beyond algorithmic computation. This leads to two profound implications. First, while it may still be possible to achieve powerful AI systems endowed with quite impressive and general abilities, AGI cannot be fully attained in computational systems that are equivalent to universal Turing machines. This limitation holds for both non-embodied and embodied Turing machines, such as robots. Second, based on the fact that only true agents can harvest the power of affordances, we conclude that only biological agents are capable of generating truly open-ended evolutionary dynamics, implying that algorithmic attempts at creating such dynamics in the field of artificial life (aLife) are doomed to fail.

Our argument proceeds as follows: In Section 2, we provide a target definition for AGI and describe some major obstacles on the way to achieve it. In Section 3, we define and contrast the notion of an agent in organismic biology and AI research. Section 4 introduces the crucial role that affordances play in AGI, while Section 5 elucidates the limitations of algorithmic agents when it comes to identifying and leveraging affordances. In Section 6, we show that our argument also applies to embodied AI agents such as robots. Section 7 presents a number of possible objections to our argument. Section 8 discusses the necessity of bio-agency for open-ended evolution. Finally, Section 9 concludes the discussion with a few remarks on the scientific and societal implications of our argument.

2. OBSTACLES TOWARD ARTIFICIAL GENERAL INTELLIGENCE

The proposal for the Dartmouth Summer Research Project begins with an ambitious statement: "An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (McCarthy et al., 1955). Over the 66 years that have passed since this was written, the field of AI research has made enormous progress, and specialized AI systems have been developed that find application across almost all aspects of human life today (see Introduction). However, the original goal of devising a system capable of integrating *all* the various capabilities required for "true machine intelligence" has not yet been reached.

According to Roitblat (2020), the defining characteristics of *general intelligence* are:

- reasoning and problem-solving,
- learning,
- inference-making,
- using common-sense knowledge,
- autonomously defining and adjusting goals,
- dealing with ambiguity and ill-defined situations, and
- creating new representations of the knowledge acquired.

Some of these capabilities are easier to formalize than others. Automated reasoning, problem-solving, learning, and inference-making, for example, can be grounded in the principles of formal logic, and are reaching impressive levels of sophistication in contemporary deep-learning approaches (Russell and Norvig, 2021). In contrast, the complete algorithmic formalization of the other items on the list remains elusive. We will discuss the problem of autonomously defining goals shortly. The three remaining characteristics are not only hard to implement algorithmically, but are difficult to define precisely in the first place. This vagueness is of a semantic and situational nature: it concerns the meaning of concepts to an agent, the knower, in their particular circumstances.

For example, we have no widely agreed-upon definition of what "common-sense knowledge" is. In fact, it is very likely that there is no generalizable definition of the term, as "common sense" represents a kind of perspectival knowing that depends radically on context. It represents a way of reacting to an everyday problem that is shared by many (or all) people at a given location and time. It is thus an intrinsically situational and normative concept, and its meaning can shift drastically across different societal and historical contexts. What it would mean for a computer to have "common sense" remains unclear: does it have to act in a way that humans of its time and location would consider commonsensical? Or does it have to develop its own kind of computer-specific, algorithmic "common sense"? What would that even mean?

Exactly the same problem affects the ability of AI algorithms to create new representations of knowledge. Those representations must not only correspond to some state of affairs in the world, but must also be relatable, understandable, and useful to some kind of knowing agent. They must represent something *to someone*. But who? Is the task of AGI to generate representations for human understanding? If not, what kind of sense does it make for a purely algorithmic system to generate representations of knowledge? It does not need them, since it does not use visualizations or metaphors for reasoning and understanding. Again, the semantic nature of the problem makes it difficult to discuss within the purely syntactic world of algorithmic AIs.

Since they cannot employ situational knowledge, and since they cannot represent and reason metaphorically, AI systems largely fail at dealing with and exploiting *ambiguities* (Byers, 2010). These limitations have been identified and formulated as *the frame problem* more than fifty years ago by Dreyfus (1965) (see also Dreyfus, 1992). Today, they are still with us as major obstacles for achieving AGI. What they have in common is an inability of algorithmic systems to reckon with the kind of uncertainty, or even paradox, that arises from context-dependent or ill-defined concepts. In contrast, the tension created by such unresolved states of knowing is often a crucial ingredient for human creativity and invention (see, for example, Scharmer and Senge, 2016).

Let us argue the case with an illustrative example. The ability to exploit ambiguities plays a role in almost any human cognitive activity. It can turn up in the most unexpected places, for instance, in one of the most rule-based human activities—an activity that we might think should be easy to formalize. As Byers beautifully observes about creativity in mathematics, "[a]mbiguity, which implies the existence of multiple, conflicting frames of reference, is the environment that gives rise to new mathematical ideas. The creativity of mathematics does not come out of algorithmic thought" (Byers, 2010, page 23). In situated problem-solving, ambiguity is oftentimes the cornerstone of a solution process. Let us consider the mathematical riddle in **Figure 1**: if we break ambiguities by taking a purely formalized algebraic perspective, the solution we find is hardly simple.[1] Yet, if we change perspective and we observe the graphical shape of the digits, we can easily note that what is summed up are the closed loops present in the numeric symbols. It turns out that the puzzle, as it is formulated, requires the ability to observe from different perspectives, to dynamically shift perceptive and

[1] In the sense of a suitable model explaining the riddle. See Burnham and Anderson (2002).

3 + 8 = 2	7 + 2 = 0	
5 + 1 = 0	6 + 6 = 2	5 + 6 = ?
9 + 1 = 1	0 + 8 = 3	

FIGURE 1 | A riddle found by one of the authors in a paper left in the coffee room of a department of mathematics.

cognitive frames, mixing both graphical and algebraic approaches for a simple solution.

Following Byers, we observe that even a strongly formalized human activity—the process of discovery in mathematics—is not entirely captured by an algorithmic search. A better metaphor would be an erratic walk across dark rooms. As Andrew Wiles describes his journey to the proof of the Fermat conjecture,[2] the solution process starts from a dark room where we "stumble around bumping into the furniture;" suddenly we find the light switch and, in the illuminated room, we "can see were we were"—an insight! Then, we move to an adjacent dark room and continue this process finding successive "light switches" into further dark rooms until the problem, at last, is solved. Each step from one room to the other is an *insight*, not a deduction and not an induction. The implication is fundamental: the mathematician comes to know a new world via an insight. The insight itself is not algorithmic. It is an act of *semantic meaning-making*. Roger Penrose makes the same point in the *Emperor's New Mind* (Penrose, 1989).

Human creativity, in all kinds of contexts, seems to require frame-switching between metaphorical or formal representations, alongside our capabilities of dealing with contradictions and ambiguities. These are not only hallmarks of human creative processes, but should also characterize AGI systems. As we will see, these abilities crucially rely on *affordances* (Gibson, 1966). Therefore, we must ask whether universal Turing machines can identify and exploit affordances. The initial step toward an answer to this question lies in the recognition that affordances arise from interactions between an agent and its umwelt. Therefore, we must first understand what an agent is, and how the concept of an "agent" is defined and used in biology and in AI research.

3. BIO-AGENCY: CONTRASTING ORGANISMS TO AI AGENTS

So far, we have avoided the question how an AGI could choose and refine its own goals (Roitblat, 2020). This problem is distinct, but still tightly related to the issues of ambiguity and representation discussed in the previous section. Selecting goals has two aspects. The first is that one must motivate the choice of a goal. One must want to reach some goal to have a goal at all, and one must have needs to want something. The other aspect is to prioritize some particular goal over a set of alternatives, according to the salience and the alignment of the chosen goal with one's own needs and capabilities in a given context.

Choosing a goal, of course, presupposes a certain *autonomy*, *i. e.*, the ability to make a "choice" (Moreno and Mossio, 2015). Here, we must emphasize again that our use of the term "choice" does *not* imply consciousness, awareness, mental states, or even cognition, which we take to involve at least some primitive kind of nervous system (Barandiaran and Moreno, 2008). It simply amounts to a system which is capable of selecting from a more or less diversified repertoire of alternative dynamic behaviors ("actions") that are at its disposal in a given situation (Walsh, 2015). All forms of life—from simple bacteria to sophisticated humans—have this capability. The most central distinction to be made here is that the selection of a specific behavior is not purely reactive, not entirely determined by environmental conditions, but (at least partially) originates from and depends on the internal organization of the system making the selection. This implies some basic kind of *agency* (Moreno and Mossio, 2015). In its broadest sense, "agency" denotes the ability of a system to initiate actions from within its own boundaries, causing effects that emanate from its own internal dynamics.

Agency requires a certain type of functional organization. More specifically, it requires *organizational closure* (Piaget, 1967; Moreno and Mossio, 2015), which leads to autopoietic (*i. e.,* self-making, self-maintaining, and self-repairing) capabilities (Maturana and Varela, 1973, 1980). It also leads to *self-determination* through self-constraint: by maintaining organizational closure, an organism is constantly providing the conditions for its own continued existence (Bickhard, 2000; Mossio and Bich, 2017). This results in the most basic, metabolic, form of autonomy (Moreno and Mossio, 2015). A minimal *autonomous agent* is a physically open, far-from-equilibrium, thermodynamic system capable of self-reproduction and self-determination.

Organisms, as autonomous agents, are *Kantian wholes*, *i. e.*, organized beings with the property that the parts exist for and by means of the whole (Kant, 1892; Kauffman, 2000, 2020). "Whole" indicates that organizational closure is a systems-level property. In physical terms, it can be formulated as a *closure of constraints* (Montévil and Mossio, 2015; Moreno and Mossio, 2015; Mossio et al., 2016). Constraints change the dynamics of the underlying processes without being altered themselves (at least not at the same time scale). Examples of constraints in organisms include enzymes, which catalyze biochemical reactions without being altered in the process, or the vascular system in vertebrates, which regulates levels of nutrients, hormones, and oxygen in

[2] *Nova* interview, https://www.pbs.org/wgbh/nova/article/andrew-wiles-fermat.

different parts of the body without changing itself at the time scale of those physiological processes (Montévil and Mossio, 2015).

It is important to note that constraint closure does not imply a fixed (static) network of processes and constraints. Instead, *organizational continuity* is maintained if the current closed organization of a system causally derives from previous instantiations of organizational closure, that is, its particular organized state at this moment in time is *dynamically presupposed* by its earlier organized states (Bickhard, 2000; DiFrisco and Mossio, 2020). Each successive state can (and indeed must) differ in their detailed physical structure from the current state. To be a Kantian whole, an autonomous system must perform at least one work-constraint cycle: it must perform physical work to continuously (re)constitute closure through new as well as recurring constraints (Kauffman, 2000, 2003; Kauffman and Clayton, 2006). Through each such cycle, a particular set of constraints is propagated, selected from a larger repertoire of possible constraints that all realize closure. In this way, the system's internal dynamics kinetically "lift" a set of mutually constituting processes from the totality of possible dynamics. This is how organizational closure leads to autopoiesis, basic autonomy, and self-determination by self-constraint: the present structure of the network of interacting processes that get "lifted" is (at least to some degree) the product of the previous unfolding of the organized network. In this way, organization maintains and propagates itself.

However, one key ingredient is still missing for an agent that actively chooses its own goals. The basic autonomous system we described above can maintain (and even repair) itself, but it cannot adapt to its circumstances—it cannot react adequately to influences from its environment. This adaptive capability is crucial for prioritizing and refining goals according to a given situation. The organism can gain some autonomy over its interactions with the environment if it is capable of regulating its own boundaries. These boundaries are required for autopoiesis, and thus must be part of the set of components that are maintained by closure (Maturana and Varela, 1980). Once boundary processes and constraints have been integrated into the closure of constraints, the organism has attained a new level of autonomy: *interactive autonomy* (Moreno and Mossio, 2015). It has now become a fully-fledged *organismal agent*, able to perceive its environment and to select from a repertoire of alternative actions when responding to environmental circumstances based on its internal organization. Expressed a bit more colloquially, making this selection requires being able to perceive the world and to evaluate "what's good or bad for me," in order to act accordingly. Here, the transition *from matter to mattering* takes place.

Interactive autonomy provides a naturalistic (and completely scientific) account of the kind of *bio-agency* (and the particular kind of goal-directedness or teleology that is associated with it, Mossio and Bich, 2017), which grounds our examination of how organisms can identify and exploit affordances in their umwelt. But before we get to this, let us contrast the complex picture of an organismal agent as a Kantian whole with the much simpler concept of an agent in AI research. In the context of AI, "[a]n agent is anything that can be viewed as *perceiving* its environment through *sensors and acting* upon that environment through *effectors*" (Russell and Norvig, 2021, original emphasis). In other words, an AI agent is an input–output processing device. Since the point of AI is to do "a good job of acting on the environment" (Russell and Norvig, 2021), the internal processing can be quite complicated, depending on the task at hand. This very broad definition of an AI agent in fact includes organismal agents, since it does not specify the kind of processes that mediate between perception and action. However, although not always explicitly stated, it is generally assumed that input-output processing is performed by some sort of algorithm that can be implemented on a universal Turing machine. The problem is that such algorithmic systems have no freedom from immediacy, since all their outputs are determined entirely—even though often in intricate and probabilistic ways—by the inputs of the system. There are no actions that emanate from the historicity of internal organization. *There is, therefore, no agency at all in an AI "agent."* What that means and why it matters for AGI and evolution will be the subject of the following sections.

4. THE KEY ROLE OF AFFORDANCES

Having outlined a suitable naturalistic account of bio-agency, we can now revisit the issue of identifying and exploiting affordances in the umwelt, or perceived environment, of an organism. The concept of an *affordance* was first proposed by Gibson (1966) in the context of ecological psychology. It was later adopted to diverse fields of investigation such as biosemiotics (Campbell et al., 2019) and robotics (Jamone et al., 2016). "Affordances" refer to what the environment offers to an agent (in the organismic sense defined above), for "good or ill." They can be manifested as opportunities or obstacles on our path to attain a goal. A recent philosophical account emphasizes the relation between the agent and its perceived environment (its umwelt), stating that affordances guide and constrain the behavior of organisms, precluding or allowing them to perform certain actions, showing them what they can and cannot do (Heras-Escribano, 2019, p. 3). A step, for instance, affords us the action of climbing; a locked door prevents us from entering. Affordances fill our world with meaning: organisms do not live in an inert environment, but "are surrounded by promises and threats" (Heras-Escribano, 2019, p. 3).

The dialectic mutual relationship between goals, actions, and affordances is of crucial importance here (Walsh, 2015). Affordances, as we have seen, require an agent with goals. Those goals motivate the agent to act. The agent first chooses which goal to pursue. It then selects an action from its repertoire (see Section 3) that it anticipates to be conducive to the attainment of the goal. This action, in turn, may alter the way the organism perceives its environment, or it may alter aspects of the environment itself, which leads to an altered set of affordances present in its umwelt. This may incite the agent to choose an alternative course of action, or even to reconsider its goals. In addition, the agent can learn to perform new actions or develop new goals along the way. This results in a constructive co-emergent dynamic in which sets of goals, actions, and affordances continuously generate and collapse each other as the world of the agent keeps entering into the next space of possibilities, its next *adjacent possible* (Kauffman, 2000). Through this co-emergent dialectic, new goals, opportunities, and ways of acting constantly arise. Since the universe is vastly non-ergodic, each moment in time provides its own unique set of opportunities and obstacles, affording new kinds of goals and actions (Kauffman, 2000). In this way, true novelty enters into the world through *radical emergence*—the generation, over time, of opportunities and rules of engagement and interaction that did not exist at any previous time in the history of the universe.

A notable example of such a co-emergent process in a human context is *jury-rigging*: given a leak in the ceiling, we cobble together a cork wrapped in a wax-soaked rag, stuff it into the hole in the ceiling, and hold it in place with duct tape (Kauffman, 2019). In general, solving a problem through jury-rigging requires several steps and involves different objects and actions, which articulate together toward a solution of the problem, mostly without any predetermined plan. Importantly, jury-rigging uses only specific subsets of the totality of *causal properties* of each object involved. Often, these properties do not coincide with previously known functional features of the object. Consider a tool, like a screwdriver, as an example. Its original purpose is to tighten screws. But it can also be used to open a can of paint, wedge a door open, scrape putty off the window, to stab or poke someone (please don't), or (should you feel so inclined) to pick your nose with it. What is important to note here is that any physical object has an *indefinite* number of alternative uses in the hands of an agent (Kauffman, 1976). This does not mean that its uses are *infinite*—even if they might be—but rather that *they cannot be known* (and thus prestated) in advance.

Ambiguity and perspective-taking also play a fundamental role in jury-rigging, as the goal of the task is to find suitable *novel* causal properties of the available objects to solve the problem at hand. The same happens in an inverse process, where we observe an artifact (or an organism Kauffman, 2019), and we aim at providing an explanation by articulating its parts, along with the particular function they carry out. For example, if we are asked what the use of an automobile is, we would probably answer that it is a vehicle equipped with an engine block, wheels, and other parts, whose diverse causal features can be articulated together to function as a locomotion and transportation system. This answer resolves most ambiguities concerning the automobile and its parts by providing a coherent frame in which the parts of the artifact are given a specific function, aimed at explaining its use as a locomotion and transportation system. In contrast, if one supposes that the purpose of an automobile is to fry eggs, one would partition the system into different sets of parts that articulate together in a distinct way such that eggs can be fried on the hot engine block. In short, for the inverse process with regard to artifacts (or organisms) what we "see it as doing" drives us to decompose the system into parts in different ways (Kauffman, 1976). Each such decomposition identifies precisely that subset of the causal properties of the identified parts that articulate together to account for and explain "what the system is doing" according to our current frame. It is critical to note that there is no universal or unique decomposition, since the way to decompose the system depends on its use and context (see also Wimsatt, 2007).

To close the loop of our argument, we note that the prospective uses of an object (and hence the decomposition we choose to analyze it) depend on the goals of the agent using it, which, in turn, depend on the agent's repertoire of actions and the affordances available to it, which change constantly and irreversibly over time. It is exactly *because* all of these are constantly evolving through their co-emergent dialectic interactions that the number of uses of an object remains *indefinite* and, in fact, *unknowable* (Kauffman, 2019). Moreover, and this is important: there is no deductive relation between the uses of an object. Take, for example, an engine block, designed to be a propulsive device in a car. It can also serve as the chassis of a tractor. Furthermore, one can use it as a bizarre (but effective) paper weight, its cylinder bores can host bottles of wine, or it can be used to crack open coconuts on one of its corners. In general, we cannot know the number of possible uses of an engine block, and we cannot deduce one use from another: the use as a paper weight abstracts from details that can conversely be necessary for it to be used to crack open coconuts. As Robert Rosen put it, complex systems invariably retain hidden properties, and their manipulation can always result in unintended consequences (Rosen, 2012). Even worse, we have seen that the relation between different uses of a thing is merely *nominal*, as there is no kind of ordering that makes it possible to relate them in a more structured way (Kauffman, 2019; Kauffman and Roli, 2021b).

This brings us to a cornerstone of our argument: when jury-rigging, it is impossible to compose any sort of well-defined list of the possible uses of the objects to be used. By analogy, **it is impossible to list all possible goals, actions, or affordances of an organismic agent in advance. In other words, Kantian wholes can not only identify and exploit affordances, but they constantly *generate* new opportunities for themselves *de novo*.** Our next question is: can algorithmic systems such as AI "agents" do this?

5. THE BOUNDED RATIONALITY OF ALGORITHMS

In the introduction, we have defined an algorithm as a computational process that can run on a universal Turing machine. This definition considers algorithms in a broad sense, including computational processes that do not halt. All algorithms operate deductively (Kripke, 2013). When implementing an algorithm as a computer program by means of some kind of formal language (including those based on recursive functional programming paradigms), we must introduce specific data and code structures, their properties and interactions, as well as the set of operations we are allowed to perform on them, in order to represent the objects and relations that are relevant for our computation. In other words, we must provide a precisely defined *ontology* on which the program can operate deductively, *e.g.,* by drawing inferences or by ordering tasks for solving a given problem. In an algorithmic framework, novelty can only be represented combinatorially: it manifests as new combinations, mergers, and relations between objects in a (potentially vast, but predefined) space of possibilities. This means that an algorithm cannot discover or generate truly novel properties or relations that were not (at least implicitly) considered in its original ontology. Therefore, an algorithm operating in a deductive manner cannot jury-rig, since it cannot find new causal properties of an object that were not already inherent in its logical premises.

To illustrate this central point, let us consider *automated planning*: a planning program is given an initial state and a predefined goal, and its task is to find a feasible—and ideally optimal—sequence of actions to reach the goal. What makes this approach successful is the possibility of describing the objects involved in the task in terms of their properties, and of representing actions in terms of the effects they produce on the world delimited by the ontology of the program, plus the requirements that need to be satisfied for their application. For the planner to work properly, there *must* be deductive relations among the different uses of an object, which are exploited by the inference engine to define an evaluation function that allows it to arrive at a solution. The problem with the planner is that, in general, there is *no deductive relation* between the possible uses of an object (see Section 4). From the use of an engine block as a paper weight, the algorithm cannot deduce its use as a method to crack open coconuts. It can, of course, find the latter use if it can be deduced, *i. e.*, if there are: *(i)* a definitive list of properties, including the fact that the engine block has rigid and sharp corners, *(ii)* a rule stating that one can break objects in the class of "breakable things" by hitting them against objects characterized by rigid and sharp corners, and *(iii)* a fact stating that coconuts are breakable.

The universe of possibilities in a computer program—however, broadly construed—is like a world of LEGO$^{\text{TM}}$ bricks: components with predefined properties and compositional relations can generate a huge space of possible combinations, even unbounded if more bricks can always be supplemented. However, if we add scotch tape, which makes it possible to assemble bricks without being constrained by their compositional mechanism, and a cutter, which enables us to cut the bricks into smaller pieces of any shape, then rules and properties are no longer predefined. We can no longer prestate a well-defined list of components, with associated properties and relations. We now have a universe of indefinite possibilities, and we are no longer trapped inside the formal frame of algorithms. *Formalization has reached its limits.* What constitutes a meaningful compositional relation becomes a semantic question, depending on our particular circumstances and the whims of our creative mind. Our possibilities may not be infinite, but they become impossible to define in advance. And because we can no longer list them, we can no longer treat them in a purely algorithmic way. This is how human creativity transcends the merely combinatorial innovative capacities of any AI we can build today. **Algorithms cannot take or shift perspective and that is why they cannot leverage ambiguity for innovation in the way an organismic agent can. Algorithms cannot jury-rig**.

At the root of this limitation is the fact that algorithms cannot want anything. To want something implies having goals that matter to us. We have argued in Section 3, that only organismic agents (but not algorithmic AI "agents") can have goals, because of their being Kantian wholes with autopoietic organization and closure of constraints. Therefore, nothing matters to an algorithm. But without mattering or goals, an algorithm has no means to identify affordances (in fact, it has no affordances), unless they are already formally predefined in its ontology, or can be derived in some logical way from predefined elements of that ontology. Thus, the algorithm cannot generate meaning where there was none before. It cannot engage in the process of *semiosis* (Peirce, 1934, p. 488). For us to make sense of the world, we must take a perspective: we must see the world from a specific point of view, contingent on our nature as fragile, limited, mortal beings which circumscribes our particular goals, abilities, and affordances. This is how organismic agents generate new frames in which to formalize posssibilities. This is how we tell what is relevant to us from what is not. Algorithms cannot do this, since they have no point of view, and require a predefined formal frame to operate deductively. To them, everything and nothing is relevant at the same time.

Now, we must draw our attention to an issue that is often neglected when discussing the nature of general intelligence: for a long time, we have believed that coming to know the world is a matter of induction, deduction, and abduction (see, for example, Hartshorne and Weiss, 1958; Mill, 1963; Ladyman, 2001; Hume, 2003; Okasha, 2016; Kennedy and Thornberg, 2018). Here, we show that this is not enough.

Consider *induction*, proceeding from a finite set of examples to an hypothesis of a universal. We observe many black ravens and formulate the hypothesis that "all ravens are black." Observe that the relevant variables and properties are already prestated, namely "ravens" and "black." Induction is over already identified features of the world and, by itself, does not identify new categories. In induction, there is an imputation of a property of the world (black) with respect to things we have already identified (ravens). There is however no insight with respect to new features of the world (*cf.* Section 2). Let us pause to think about this: induction by itself cannot reveal novel features of the world—features that are not already in our ontology.

This is even more evident for *deduction*, which proceeds from prestated universal categories to the specific. "All men are mortal, Socrates is a man, therefore Socrates is a mortal." All theorems and proofs in mathematics have this deductive structure. However, neither induction nor deduction by themselves can reveal novel features of the world not already in our ontology.

Finally, we come to *abduction*, which aims at providing an explanation of an observation by asserting an already known precondition that is likely to have this observation as a consequence. For example, if we identify an automobile as a means of locomotion and transportation, and had decomposed it into parts that articulate together to support its function as a means of locomotion and transportation, we are then able to explain its failure to function in this sense by a failure of one of its now defined parts. If the car does not turn on, we can suppose the battery is dead. Abduction is differential diagnosis from a prestated set of conditions and possibilities that articulate to carry out what we "see the system as doing or being." But there is no unique decomposition. The number of decompositions is indefinite. Therefore, when implemented in a computer program, this kind of reasoning cannot reveal novel features of the world not already in the ontology of the program.

To summarize: with respect to coming to know the world, **once we have carved the world into a finite set of categories, we can no longer see the world beyond those categories**. In other words, new meanings—along with their symbolic grounding in real objects—are outside of the predefined ontology of an agential system. The same limitation also holds for probabilistic forms of inference, involving, *e.g.,* Bayesian nets (see Gelman et al., 2013). Consider the use of an engine block as a paper weight, and a Bayesian algorithm updating to improve engine blocks with respect to functioning as a paper weight. No such updating will reveal that engine blocks can also be used to crack open coconuts. The priors for such an innovation could not be deduced, even in principle. Similarly, Markov blankets (see, for example, Hipólito et al., 2021) are restricted to pre-existing categories.

Organisms come to know new features of the world by semiosis—a process which involves *semantic meaning-making* of the kind described above, not just formal (syntactic) reasoning through deduction, induction, or abduction. This is true of mathematicians. It is also true of Caledonian crows who solve problems of astonishing complexity, requiring sophisticated multi-step jury-rigging (Taylor et al., 2010). Chimpanzees learning to use tools have the same capacity to improvise (Köhler, 2013). Simpler organisms—down to bacteria—must have it too, although probably in a much more limited sense. After all, they are at the basis of an evolutionary process toward more complex behavior, which presupposes the identification and exploitation of new opportunities. Our human ontology has evolved into a much more complex state than that of a primitive unicellular organism. In general, all organisms act in alignment with their goals, capabilities, and affordances (see Section 4), and their agential behavior can undergo variation and selection. A useful action—exploiting a novel affordance—can be captured by heritable variation (at the genetic, epigenetic, behavioral, or cultural level) and thus passed on across generations. This "coming to know the world" is what makes the evolutionary expansion of our ontologies possible. It goes beyond induction, deduction, and abduction. Organisms can do it, but universal Turing machines cannot.

In conclusion, the rationality of algorithms is bounded by their ontology. However vast this ontology may be, algorithms cannot transcend their predefined limitations, while organisms can. This leads us to our central conclusion, which is both radical and profound: **not all possible behaviors of an organismic agent can be formalized and performed by an algorithm—not all organismic behaviors are Turing-computable. Therefore, organisms are not Turing machines. It also means that true AGI cannot be achieved in an algorithmic frame**, since AI "agents" cannot choose and define their own goals, and hence exploit affordances, deal with ambiguity, or shift frames in ways organismic agents can. Because of these limitations, algorithms cannot evolve in truly novel directions (see Section 8 below).

6. IMPLICATIONS FOR ROBOTS

So far, we have only considered algorithms that run within some stationary computing environment. The digital and purely virtual nature of this environment implies that all features within in must, by definition, be formally predefined. Its digital environment, in its finite totality, *is* the ontology of an AI algorithm. There is nothing outside it for the AI "agent" to discover. The real world is not like that. We have argued in the previous sections that our world is full of surprises that cannot be entirely formalized, since not all future possibilities can be prestated. Therefore, the question arises whether an AI agent that does get exposed to the real world could identify and leverage affordances when it encounters them.

In other words, does our argument apply to *embodied Turing machines*, such as robots, that interact with the physical world through sensors and actuators and may be able to modify their bodily configuration? The crucial difference to a purely virtual AI "agent" is that the behavior of a robot results from interactions between its control program (an algorithm), its physical characteristics (which define its repertoire of actions), and the physical environment in which it finds itself (Pfeifer and Bongard, 2006). Moreover, learning techniques are put to powerful use in robotics, meaning that robots can adapt their behavior and improve their performance based on their relations to their physical environment. Therefore, we can say that robots are able to learn from experience and to identify specific sensory-motor patterns in the real world that are useful to attain their goals (Pfeifer and Scheier, 2001). For instance, a quadruped robot controlled by an artificial neural network can learn to control its legs on the basis of the forces perceived from the ground, so as to develop a fast and robust gait. This learning process can be guided either by a task-oriented evaluation function, such as forward gait speed, or a task-agnostic one that rewards coordinated behaviors (Prokopenko, 2013), or both.

Does that mean that robots, as embodied Turing machines, can identify and exploit affordances? Does it mean that robots, just like organisms, have an umwelt full of opportunities and threats? As in the case of stationary AI "agents," the answer is a clear and resounding "no." The same problems we have discussed in the previous sections also affect robotics. Specifically, they manifest themselves as the symbol grounding problem and the frame problem. The *symbol grounding problem* concerns the issue of attaching symbols to sensory-motor patterns (Harnad, 1990). It amounts to the question whether it is feasible for a robot to detect relevant sensory-motor patterns that need to be associated with new concepts—*i. e.*, new variables in the ontology of the robot. This, in turn, leads to the more general *frame problem* (see Section 2 and McCarthy and Hayes, 1969): the problem of specifying in a given situation what is relevant for a robot's goals. Again, we run into the problem of choosing one's own goals, of shifting frames, and of dealing with ambiguous information that cannot be formalized in the form of a predefined set of possibilities.

As an example, consider the case of a robot whose goal it is to open coconuts. Its only available tool is an engine block, which it currently uses as a paper weight. There are no other tools, and the coconuts cannot be broken by simply throwing them against a wall. In order to achieve its goal, the robot must acquire information on the relevant causal features of the engine block to open coconuts. Can it exploit this affordance? The robot can move around and perceive the world via its sensors. It can acquire experience by performing random moves, one of which may cause it to hit the engine block, to discover that the block has the property of being "hard and sharp," which is useful for cracking the nut. However, how does the robot know that it needs to look for this property in the objects of its environment? This is but the first useful step in solving the problem. By the same random moves, the robot might move the engine block, or tip it on its side. How can the robot "understand" that "hard and sharp" will prove to be useful, but "move to the left" will not? How long will this single step take?

Furthermore, if the coconut is lying beside the engine block, tipping it over may lead to the nut being cracked as well. How can the robot connect several coordinated causal features to achieve its goal, if none of them can be deduced from the others? The answer is: it cannot. We observe that **achieving the final goal may require connecting several relevant coordinated causal features of real-world objects, none of which is deducible from the others**. This is analogous to the discovery process in mathematics we have described in Section 2: wandering through a succession of dark rooms, each transition illuminated by the next in a succession of insights. There is no way for the robot to know that it is improving over the incremental steps of its search. Once an affordance is identified, new affordances emerge as a consequence and the robot cannot "know" in advance that it is accumulating successes until it happens upon the final achievement: there is no function optimization to be performed over such a sequence of steps, no landscape to search by exploiting its gradients, because each step is a search in a space of possibilities that cannot be predefined. The journey from taking the first step to reaching the ultimate goal is blind luck over some unknown time scale. With more steps, it becomes increasingly difficult to know if the robot improves, since reaching the final goal is in general not an incremental process.

The only way to achieve the robot's ultimate goal is for it to already have a preprogrammed ontology that allows for multi-step inferences. Whether embodied or not, the robot's control algorithm can only operate deductively. But if the opportunity to crack open coconuts on the engine block has been predefined, then it does not really count as discovering a new causal property. It does not count as exploiting a novel affordance. **Robots do not generate new opportunities for themselves in the way organisms do. Even though engaging with their environment, they cannot participate in the emergent triad of goals, actions, and affordances** (see Section 4). Therefore, we must conclude that its embodied nature does not really help a robotic algorithm to achieve anything resembling true AGI.

7. POSSIBLE OBJECTIONS

We suspect that our argument may raise a number of objections. In this section, we anticipate some of these, and attempt to provide adequate replies.

A first potential objection concerns *the ability of deep-learning algorithms to detect novel correlations* in large data sets in an apparently hypothesis-free and unbiased manner. The underlying methods are mainly based on complex network models, rather than traditional sequential formal logic. When the machine is trained with suitable data, shouldn't it be able to add new symbols to its ontology that represent the newly discovered correlations? Would this not count as identifying and exploiting a new affordance? While it is true that the ontology of such a deep-learning machine is not explicitly predefined, it is nevertheless implicitly given through the constraints of the algorithm and the training scenario. Correlations can only be detected between variables that are defined through an external model of the data. Moreover, all current learning techniques rely on the maximization (or minimization) of one or more evaluation functions. These functions must be provided by the designers of the training scenario, who thus determine the criteria for performance improvement. The program itself does not have the ability of choosing the goal of the task at hand. This even holds for task-agnostic functions of learning scenarios, as they again are the result of an imposed external choice. In the end, with no bias or hypothesis at all, what should the learning program look for? In a truly bias- or hypothesis-free scenario (if that is possible at all), any regularity (even if purely accidental) would become meaningful (Calude and Longo, 2017), which results in no meaning at all. Without any goal or perspective, there is no insight to be gained.

A second objection might be raised concerning the rather common observation that AI systems, such as programs playing chess or composing music, often surprise us or behave in unpredictable ways. However, *machine unpredictability* does not imply that their behavior is not deducible. Instead, it simply means that we cannot find an explanation for it, maybe due to a lack of information, or due to our own limited cognitive and/or computational resources. For example, a machine playing chess can take decisions by exploiting a huge repertoire of moves, and this may produce surprising behavior in the eye of the human opponent, since it goes far beyond our own cognitive capacity. Nevertheless, the behavior of the machine is deductively determined, ultimately based on simple combinatorics. More generally, it is well-known that there are computer programs whose output is not compressible. Their behavior cannot be predicted other than actually running the full program. This computationally irreducible behavior cannot be anticipated, but it is certainly algorithmic. Due to their competitive advantage when dealing with many factors, or many steps, in a deductive procedure, AI "agents" can easily fool us by mimicking creative behavior, even though their algorithmic operation does not allow for the kind of semantic innovation even a simple organism is capable of.

A third objection could be that our argument carelessly ignores potential progress in computational paradigms and robot design that may lead to a solution of the apparently irresolvable problems we present here. A common futurist scenario in this context is one in which AI "agents" themselves replace human engineers in designing AI architectures, leading to *a technological singularity*—a technology which is far beyond human grasp (see, for example, Vinge, 1993; Kurzweil, 2005; Eden et al., 2013; Bostrom, 2014; Shanahan, 2015; Chalmers, 2016). We are sympathetic to this objection (although not to the notion of a singularity based on simple extrapolation of our current capabilities). Our philosophical approach is exactly based on the premise that the future is open, and will always surprise us in fundamentally unpredictable ways. But there is no paradox here: what we are arguing for is that AGI is impossible *within the current algorithmic frame* of AI research, which is based on Turing machines. We are open to suggestions how the limitations of this frame could be transcended. One obvious way to do this is a biological kind of robotics, which uses organismic agents (such as biological cells) to build organic computation devices or robots. We are curious (and also apprehensive) concerning the potential (and dangers) which such non-algorithmic frameworks hold for the future. An AGI which *could* indeed choose its own goals, would not be aligned with our own interests (by definition), and may not be controllable by humans, which seems to us to defy the purpose of generating AI as a benign and beneficial technology in the first place.

One final, and quite serious, philosophical objection to our argument is that it may be impossible to empirically distinguish between *a sophisticated algorithm mimicking agential behavior*, and true organismic agency as outlined in Section 3. In this case, our argument may be of no practical importance. It is true that humans are easily fooled into interpreting completely mechanistic behavior in intentional and teleological terms. Douglas Hofstadter (2007), for example, mentions a dot of red light that is moving along the walls of the San Francisco Exploratorium, responding by simple feedback to movements of the museum visitors. Every time a visitor tries to touch the dot, it seems to escape at the very last moment. Even though based on a simple feedback mechanism, it is tempting to interpret such behavior as intentional.[3] Could we have fallen prey to such an illusion when interpreting the behavior of organisms as true agency? We do not think so. First, the organizational account of agency we rely on not only accounts for goal-oriented behavior, but also for basic functional properties of living systems, such as their autopoietic ability to self-maintain and self-repair. Thus, agency is a higher-level consequence of more basic abilities of organisms that cannot easily be accounted for by alternative explanations. Even though these basic abilities have not yet been put to the test in a laboratory, there is no reason to think that they won't be in the not-too-far future. Second, we think the account of organismic agency presented here is preferable over an algorithmic explanation of "agency" as evolved input-output processing, since it has much greater explanatory power. It takes the phenomenon of agency seriously instead of trying to explain it away. Without this conceptual framework, we could not even ask the kind of questions raised in this paper, since they would never arise within an algorithmic framework. In essence, the non-reductionist (yet still naturalist) world we operate in is richer than the reductionist one in that it allows us to deal scientifically with a larger range of undoubtedly interesting and relevant phenomena (see also Wimsatt, 2007).

8. OPEN-ENDED EVOLUTION IN COMPUTER SIMULATIONS

Before we conclude our argument, we would like to consider its implications beyond AGI, in particular, for the theory of evolution, and for research in the field of artificial life (ALife). One of the authors has argued earlier that evolvability and agency *must* go together, because the kind of organizational continuity that turns a cell cycle into a *reproducer*—the minimal unit of Darwinian evolution—also provides the evolving organism with the ability to act autonomously (Jaeger, 2022). Here, we go one step further and suggest that **organismic agency is a fundamental prerequisite for open-ended evolution**, since it enables organisms to identify and exploit affordances in their umwelt, or perceived environment. Without agency, there is no co-emergent dialectic between organisms' goals, actions, and affordances (see Section 5). And without this kind of dialectic, evolution cannot transcend its predetermined space of possibilities. It cannot enter into the next adjacent possible. It cannot truly innovate, remaining caught in a deductive ontological frame (Fernando et al., 2011; Bersini, 2012; Roli and Kauffman, 2020).

Let us illustrate this with the example of ALife. The ambitious goal of this research field is to create models of digital "organisms" that are able to evolve and innovate in ways equivalent to natural evolution. Over the past decades, numerous attempts have been made to generate open-ended evolutionary dynamics in simulations such as Tierra (Ray, 1992) and Avida (Adami and Brown, 1994). In the latter case, the evolving "organisms" reach an impressive level of sophistication (see, for example, Lenski et al., 1999, 2003; Zaman et al., 2014). They have an internal "metabolism" that processes nutrients to gain energy from their environment in order to survive and reproduce. However, this "metabolism" does not exhibit organizational closure, or any other form of true agency, since it remains purely algorithmic. And so, no matter how complicated, such evolutionary simulations always tend to get stuck at a certain level of complexity (Bedau et al., 2000; Standish, 2003). Even though some complexification of ecological interactions (*e. g.,* mimics of trophic levels or parasitism) can occur, we never observe any innovation that goes beyond what was implicitly considered in the premises of the simulation. This has led to some consternation and the conclusion that the strong program of ALife—to generate any truly life-like processes in a computer simulation—has failed to achieve its goal so far. In fact, we would claim that this failure is comprehensive: it affects all

[3] Regarding machine intentionality see also the work by Braitenberg (1986).

attempts at evolutionary simulation that have been undertaken so far. Why is that so?

Our argument provides a possible explanation for the failure of strong ALife: even though the digital creatures of Avida, for example, can exploit "new" nutrient sources, they can only do so because these sources have been endowed with the property of being a potential food source at the time the simulation was set up. They were part of its initial ontology. The algorithm cannot do anything it was not (implicitly) set up to do. Avida's digital "life forms" can explore their astonishingly rich and large space of possibilities combinatorially. This is what allows them, for example, to feed off other "life forms" to become predators or parasites. The resulting outcomes may even be completely unexpected to an outside observer with insufficient information and/or cognitive capacity (see Section 7). However, Avida's "life forms" can never discover or exploit any truly new opportunities, like even the most primitive natural organisms can. They cannot generate new meaning that was not already programmed into their ontology. They cannot engage in semiosis. What we end up is a very high-dimensional probabilistic combinatorial search. Evolution has often been likened to such intricate search strategies, but our view suggests that organismic agency pushes it beyond.

Organismic open-ended evolution into the adjacent possible requires the identification and leveraging of novel affordances. In this sense, it cannot be entirely formalized. In contrast, algorithmic evolutionary simulations will forever be constrained by their predefined formal ontologies. They will never be able to produce any true novelty, or radical emergence. They are simply not like organismic evolution since they lack its fundamental creativity. As some of us have argued elsewhere: emergence is not engineering (Kauffman and Roli, 2021a). The biosphere is an endlessly propagating adapting construction, not an entailed algorithmic deduction (Kauffman, 2019). In other words, the world is not a theorem (Kauffman and Roli, 2021b), but a neverending exploratory process. It will never cease to fascinate and surprise us.

9. CONCLUSION

In this paper, we have argued two main points: (1) AGI is impossible in the current algorithmic frame of research in AI and robotics, since algorithms cannot identify and exploit new affordances. (2) As a direct corollary, truly open-ended evolution into the adjacent possible is impossible in algorithmic systems, since they cannot transcend their predefined space of possibilities.

Our way of arriving at these conclusions is not the only possible one. In fact, the claim that organismic behavior is not entirely algorithmic was made by Robert Rosen as early as the 1950s (Rosen, 1958a,b, 1959, 1972). His argument is based on category theory and neatly complements our way of reasoning, corroborating our insight. It is summarized in Rosen's book "Life Itself" (Rosen, 1991). As a proof of principle, he devised a diagram of compositional mappings that exhibit *closure to efficient causation*, which is equivalent to organizational closure (see Section 3). He saw this diagram as a highly abstract relational representation of the processes that constitute a living system. Rosen was able to prove mathematically that this type of organization "has no largest model" (Rosen, 1991). This has often been confounded with the claim that it cannot be simulated in a computer at all. However, Rosen is not saying that we cannot generate algorithmic models of some (maybe even most) of the behaviors that a living system can exhibit. In fact, it has been shown that his diagram *can* be modeled in this way using a recursive functional programming paradigm (Mossio et al., 2009). What Rosen *is* saying is exactly what we are arguing here: there will always be some organismic behaviors that cannot be captured by a preexisting formal model. This is an *incompleteness argument* of the kind Gödel made in mathematics (Nagel and Newman, 2001): for most problems, it is still completely fine to use number theory after Gödel's proof. In fact, relevant statements about numbers that do not fit the theory are exceedingly rare in practice. Analogously, we can still use algorithms implemented by computer programs to study many aspects of organismic dynamics, or to engineer (more or less) target-specific AIs. Furthermore, it is always possible to extend the existing formal model to accommodate a new statement or behavior that does not yet fit in. However, this process is infinite. We will never arrive at a formal model that captures *all* possibilities. Here, we show that this is because those possibilities cannot be precisely prestated and defined in advance.

Another approach that comes to very similar insights to ours is *biosemiotics* (see, for example, Hoffmeyer, 1993; Barbieri, 2007; Favareau, 2010; Henning and Scarfe, 2013). Rather than a particular field of inquiry, biosemiotics sees itself as a broad and original perspective on life and its evolution. It is formulated in terms of the production, exchange, and interpretation of signs in biological systems. The process of meaning-making (or semiosis) is central to biosemiotics (Peirce, 1934). Here, we link this process to autopoiesis (Varela et al., 1974; Maturana and Varela, 1980) and the organizational account, which sees bio-agency grounded in a closure of constraints within living systems (Montévil and Mossio, 2015; Moreno and Mossio, 2015; Mossio et al., 2016), and the consequent co-emergent evolutionary dialectic of goals, actions, and affordances (Walsh, 2015; Jaeger, 2022). Our argument suggests that the openness of semiotic evolution is grounded in our fundamental inability to formalize and prestate possibilities for evolutionary and cognitive innovation in advance.

Our insights put rather stringent limitations on what traditional mechanistic science and engineering can understand and achieve when it comes to agency and evolutionary innovation. This affects the study of any kind of agential system—in computer science, biology, and the social sciences—including higher-level systems that contain agents, such as ecosystems or the economy. In these areas of investigation, any purely formal approach will remain forever incomplete. This has important repercussions for the philosophy of science: the basic problem is that, with respect to coming to know the world, once we have carved it into a finite set of categories, we can no longer see beyond those categories. The grounding of meaning in real objects is outside any predefined formal ontology. The evolution of scientific knowledge itself is entailed by no law. It cannot be formalized (Kauffman and Roli, 2021a,b).

What would such a *meta-mechanistic science* look like? This is not entirely clear yet. Its methods and concepts are only now being elaborated (see, for example, Henning and Scarfe, 2013). But one thing seems certain: it will be a science that takes agency seriously. It will allow the kind of teleological behavior that is rooted in the self-referential closure of organization in living systems. It is naturalistic but not reductive. Goals, actions, and affordances are emergent properties of the relationship between organismal agents and their umwelt—the world of meaning they live in. This emergence is of a radical nature, forever pushing beyond predetermined ontologies into the adjacent possible. This results in a worldview that closely resembles Alfred North Whitehead's *philosophy of organism* (Whitehead, 1929). It sees the world less as a clockwork, and more like an evolving ecosystem, a creative process centered around harvesting new affordances.

It should be fairly obvious by now that our argument heavily relies on teleological explanations, necessitated by the goal-oriented behavior of the organism. This may seem problematic: teleological explanations have been traditionally banned from evolutionary biology because they seemingly require (1) an inversion of the flow from cause to effect, (2) intentionality, and (3) a kind of normativity, which disqualify them from being proper naturalistic scientific explanations.

Here, we follow Walsh (2015), who provides a very convincing argument that this is not the case. First, it is important to note that we are not postulating any large-scale teleology in evolution—no omega point toward which evolution may be headed. On the contrary, our argument for open-endedness explicitly precludes such a possibility, even in principle (see Section 8). Second, the kind of teleological explanation we propose here for the behavior of organisms and its evolution is *not* a kind of causal explanation. While causal explanations state which effect follows which cause, teleological explanation deals with the conditions that are conducive for an organism to attain its goal. The goal does not *cause* these conditions, but rather presupposes them. Because of this, there is no inversion of causal flow. Finally, the kind of goal-directed behavior enabled by bio-agency does *not* require awareness, intentionality, or even cognition. It can be achieved by the simplest organisms (such as bacteria), simply due to the fact that they exhibit an internal organization based on a closure of constraints (see Section 3). This also naturalizes the kind of normativity we require for teleology (Mossio and Bich, 2017): the organism really does have a goal from which it can deviate. That goal is to stay alive, reproduce, and flourish. All of this means that there is nothing supranatural or unscientific about the kind of teleological explanations that are used in our argument. They are perfectly valid explanations. There is no need to restrict ourselves to strictly mechanistic arguments, which yield an impoverished world view since they cannot capture the deep problems and rich phenomena we have been discussing throughout this paper.

While such metaphysical and epistemological considerations are important for understanding ourselves and our place in the world, our argument also has eminently practical consequences. The achievement of AGI is often listed as one of the most threatening existential risks to the future of humanity (see, for example, Yudkowsky, 2008; Ord, 2020). Our analysis suggests that such fears are greatly exaggerated. No machine will want to replace us, since no machine will want anything, at least not in the current algorithmic frame of defining a machine. This, of course, does not prevent AI systems and robots from being harmful. Protocols and regulations for AI applications are urgent and necessary. But AGI is not around the corner, and we are not alone with this assessment. The limits of current AI applications have been questioned by others, emphasizing that these systems lack autonomy and understanding capabilities, which we conversely find in natural intelligence (Nguyen et al., 2015; Broussard, 2018; Hosni and Vulpiani, 2018; Marcus and Davis, 2019; Mitchell, 2019; Roitblat, 2020; Sanjuán, 2021; Schneier, 2021). The true danger of AI lies in the social changes and the disenfranchisement of our own agency that we are currently effecting through target-specific algorithms. It is not Skynet, but Facebook, that will probably kill us in the end.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

All authors contributed equally to this manuscript, conceived the argument, and wrote the paper together.

ACKNOWLEDGMENTS

JJ has profited from numerous discussions on the topic of organismic agency with Jan-Hendirk Hofmeyr, Matteo Mossio, and Denis Walsh, and on possibility spaces with Andrea Loettgers and Tarja Knuuttila, none of whom necessarily share the opinions expressed in this paper. JJ thanks the late Brian Goodwin for crucial early influences on his thinking. No doubt, Brian would have loved this paper.

REFERENCES

Adami, C., and Brown, C. T. (1994). "Evolutionary learning in the 2D artificial life system 'Avida'," in *Artificial Life IV: Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems*, eds P. Maes and R. Brooks (Cambridge, MA: MIT Press), 377–381.

Arnellos, A., and Moreno, A. (2015). Multicellular agency: an organizational view. *Biol. Philosophy* 30, 333–357. doi: 10.1007/s10539-015-9484-0

Arnellos, A., Spyrou, T., and Darzentas, J. (2010). Towards the naturalization of agency based on an interactivist account of autonomy. *New Ideas Psychol.* 28, 296–311. doi: 10.1016/j.newideapsych.2009.09.005

Barandiaran, X., and Moreno, A. (2008). On the nature of neural information: a critique of the received view 50 years later. *Neurocomputing* 71, 681–692. doi: 10.1016/j.neucom.2007.09.014

Barandiaran, X. E., Di Paolo, E., and Rohde, M. (2009). Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action. *Adapt. Behav.* 17, 367–386. doi: 10.1177/1059712309343819

Barbieri, M., editor (2007). *Introduction to Biosemiotics: The New Biological Synthesis*. Dordrecht, NL: Springer.

Bedau, M. A., McCaskill, J. S., Packard, N. H., Rasmussen, S., Adami, C., Green, D. G., et al. (2000). Open problems in artificial life. *Artif. Life* 6, 363–376. doi: 10.1162/106454600300103683

Bersini, H. (2012). Emergent phenomena belong only to biology. *Synthese* 185, 257–272. doi: 10.1007/s11229-010-9724-4

Bickhard, M. H. (2000). Autonomy, function, and representation. *Commun. Cogn. Artif. Intell.* 17, 111–131.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Braitenberg, V. (1986). *Vehicles: Experiments in Synthetic Psychology*. Cambridge, MA: MIT Press.

Briot, J.-P., and Pachet, F. (2020). Deep learning for music generation: challenges and directions. *Neural Comput. Appl.* 32, 981–993. doi: 10.1007/978-3-319-70163-9

Broussard, M. (2018). *Artificial Unintelligence: How Computers Misunderstand the World*. Cambridge, MA: MIT Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *arXiv preprint* arXiv:2005.14165.

Burnham, K., and Anderson, D. (2002). *Model Selection and Multi-Model Inference*, 2nd Edn, New York, NY: Springer.

Byers, W. (2010). *How Mathematicians Think*. Princeton, NJ: Princeton University Press.

Calude, C., and Longo, G. (2017). The deluge of spurious correlations in big data. *Found. Sci.* 22, 595–612. doi: 10.1007/s10699-016-9489-4

Campbell, C., Olteanu, A., and Kull, K. (2019). Learning and knowing as semiosis: extending the conceptual apparatus of semiotics. *Sign Syst. Stud.* 47, 352–381. doi: 10.12697/SSS.2019.47.3-4.01

Campbell, R. (2010). The emergence of action. *New Ideas Psychol.* 28, 283–295. doi: 10.1016/j.newideapsych.2009.09.004

Chalmers, D. (2020). GPT-3 and general intelligence. *Daily Nous* 30.

Chalmers, D. J. (2016). "The singularity: a philosophical analysis," in *Science Fiction and Philosophy*, ed S. Schneider (Hoboken, NJ: John Wiley & Sons, Inc), 171–224.

DiFrisco, J., and Mossio, M. (2020). Diachronic identity in complex life cycles: an organizational perspective," in *Biological Identity: Perspectives from Metaphysics and the Philosophy of Biology*, eds A. S. Meincke, and J. Dupré (London: Routledge).

Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York, NY: Basic Books.

Douglas Hofstadter, R. (2007). *I Am a Strange Loop*. New York, NY: Basic Books.

Dreyfus, H. (1965). *Alchemy and Artificial Intelligence*. Technical Report, RAND Corporation, Santa Monica, CA, USA.

Dreyfus, H. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press.

Eden, A. H., Moor, J. H., Soraker, J. H., and Steinhart, E., editors (2013). *Singularity Hypotheses: A Scientific and Philosophical Assessment*. New York, NY: Springer.

Favareau, D., editor (2010). *Essential Readings in Biosemiotics*. Springer, Dordrecht, NL.

Fernando, C., Kampis, G., and Szathmáry, E. (2011). Evolvability of natural and artificial systems. *Proc. Compu. Sci.* 7, 73–76. doi: 10.1016/j.procs.2011.12.023

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., and Vehtari, A. (2013). *Bayesian Data Analysis*, 3rd Edn. Boca Raton, FL: Taylor & Francis Ltd.

Gibson, J. (1966). *The Senses Considered as Perceptual Systems*. London: Houghton Mifflin.

Gold, J. I., and Shadlen, M. N. (2007). The neural basis of decision making. *Ann. Rev. Neurosci.* 30, 535–574. doi: 10.1146/annurev.neuro.29.051605.113038

Harnad, S. (1990). The symbol grounding problem. *Physica D Nonlin. Phenomena* 42, 335–346. doi: 10.1016/0167-2789(90)90087-6

Hartshorne, C., and Weiss (1958). *Collected Papers of Charles Sanders Peirce*. Boston, MA: Belknap Press of Harvard University Press.

Henning, B., and Scarfe, A., editors (2013). *Beyond Mechanism: Putting Life Back Into Biology*. Plymouth: Lexington Books.

Heras-Escribano, M. (2019). *The Philosophy of Affordances*. London: Springer.

Hipólito, I., Ramstead, M., Convertino, L., Bhat, A., Friston, K., and Parr, T. (2021). Markov blankets in the brain. *Neurosci. Biobehav. Rev.* 125, 88–97. doi: 10.1016/j.neubiorev.2021.02.003

Hoffmeyer, J. (1993). *Signs of Meaning in the Universe*. Bloomington, IN: Indiana University Press.

Hong, J.-W., and Curran, N. (2019). Artificial intelligence, artists, and art: attitudes toward artwork produced by humans vs. artificial intelligence. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* 15, 1–16. doi: 10.1145/3326337

Hosni, H., and Vulpiani, A. (2018). Data science and the art of modelling. *Lettera Matematica* 6, 121–129. doi: 10.1007/s40329-018-0225-5

Hume, D. (2003). *A Treatise of Human Nature*. Chelmsford, MA: Courier Corporation.

Jaeger, J. (2022). "The fourth perspective: evolution and organismal agency," in *Organization in Biology*, ed M. Mossio (Berlin: Springer).

Jamone, L., Ugur, E., Cangelosi, A., Fadiga, L., Bernardino, A., Piater, et al. (2016). Affordances in psychology, neuroscience, and robotics: a survey. *IEEE Trans. Cogn. Develop. Syst.* 10, 4–25. doi: 10.1109/TCDS.2016.2594134

Kant, I. (1892). *Critique of Judgement*. New York, NY: Macmillan.

Kauffman, S. (1976). "Articulation of parts explanation in biology and the rational search for them," in *Topics in the Philosophy of Biology*, (Dordrecht: Springer), 245–263.

Kauffman, S. (2000). *Investigations*. Oxford: Oxford University Press.

Kauffman, S. (2003). Molecular autonomous agents. *Philosoph. Trans. Roy. Soc. London Series A Math. Phys. Eng. Sci.* 361, 1089–1099. doi: 10.1098/rsta.2003.1186

Kauffman, S. (2019). *A World Beyond Physics: the Emergence and Evolution of Life*. Oxford: Oxford University Press.

Kauffman, S. (2020). Eros and logos. *Angelaki* 25, 9–23. doi: 10.1080/0969725X.2020.1754011

Kauffman, S., and Clayton, P. (2006). On emergence, agency, and organization. *Biol. Philosophy* 21, 501–521. doi: 10.1007/s10539-005-9003-9

Kauffman, S., and Roli, A. (2021a). The third transition in science: beyond Newton and quantum mechanics – a statistical mechanics of emergence. *arXiv preprint* arXiv:2106.15271.

Kauffman, S., and Roli, A. (2021b). The world is not a theorem. *Entropy* 23:1467. doi: 10.3390/e23111467

Kennedy, B., and Thornberg, R. (2018). "Deduction, induction, and abduction," in *The SAGE Handbook of Qualitative Data Collection* (London: SAGE Publications), 49–64.

Köhler, W. (2013). *The Mentality of Apes*. London: Routledge.

Kripke, S. (2013). "The Church-Turing "thesis" as a special corollary of Gödel's completeness theorem," in *Computability: Gödel, Turing, Church, and Beyond*, eds B. Copeland, C. Posy, and O. Shagrir (Cambridge, MA: The MIT Press), Ch. 4, 77–104.

Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology*. New York, NY: The Viking Press.

Ladyman, J. (2001). *Understanding Philosophy of Science*. London: Routledge.

LaValle, S. (2006). *Planning Algorithms*. Cambridge: Cambridge University Press.

Lenski, R. E., Ofria, C., Collier, T. C., and Adami, C. (1999). Genome complexity, robustness and genetic interactions in digital organisms. *Nature* 400, 661–664. doi: 10.1038/23245

Lenski, R. E., Ofria, C., Pennock, R. T., and Adami, C. (2003). The evolutionary origin of complex features. *Nature* 423, 139–144. doi: 10.1038/nature01568

Marcus, G. and Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York, NY: Vintage.

Marcus, G. and Davis, E. (2020). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *Technol. Rev.*

Maturana, H., and Varela, F. (1973). *De Maquinas y Seres Vivos*. Santiago: Editorial Universitaria.

Maturana, H., and Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht: Springer.

McCarthy, J., and Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Mach. Intell.* 463–502.

McCarthy, J., Minsky, M., Rochester, N., and Shannon, C. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. Available online at: http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

McShea, D. W. (2012). Upper-directed systems: a new approach to teleology in biology. *Biol. Philosophy* 27, 63–684. doi: 10.1007/s10539-012-9326-2

McShea, D. W. (2013). Machine wanting. *Stud. History Philosophy Sci. Part C Biol. Biomed. Sci.* 44, 679–687. doi: 10.1016/j.shpsc.2013.05.015

McShea, D. W. (2016). Freedom and purpose in biology. *Stud. History Philosophy Sci. Part C Biol. Biomed. Sci.* 58, 64–72. doi: 10.1016/j.shpsc.2015.12.002

Meincke, A. S. (2018). "Bio-agency and the possibility of artificial agents," in *Philosophy of Science (European Studies in Philosophy of Science), Vol. 9*, Eds. A. Christian, D. Hommen, N. Retzlaff, and G. Schurz (Cham: Springer International Publishing), 65–93.

Mill, J. (1963). *Collected Works*. Toronto, ON: University of Toronto Press.

Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. London: Penguin UK.

Montévil, M., and Mossio, M. (2015). Biological organisation as closure of constraints. *J. Theor. Biol.* 372, 179–191. doi: 10.1016/j.jtbi.2015.02.029

Moreno, A., and Etxeberria, A. (2005). Agency in natural and artificial systems. *Artif. Life* 11, 161–175. doi: 10.1162/1064546053278919

Moreno, A., and Mossio, M. (2015). *Biological Autonomy*. Dordrecht: Springer.

Mossio, M., and Bich, L. (2017). What makes biological organisation teleological? *Synthese* 194, 1089–1114. doi: 10.1007/s11229-014-0594-z

Mossio, M., Longo, G., and Stewart, J. (2009). A computable expression of closure to efficient causation. *J. Theor. Biol.* 257, 489–498. doi: 10.1016/j.jtbi.2008.12.012

Mossio, M., Montévil, M., and Longo, G. (2016). Theoretical principles for biology: organization. *Progr. Biophys. Mol. Biol.* 122, 24–35. doi: 10.1016/j.pbiomolbio.2016.07.005

Müller, V. C., and Bostrom, N. (2016). "Future progress in artificial intelligence: a survey of expert opinion," in *Fundamental Issues of Artificial Intelligence*, Vol. 376, eds V. C. Müller (Cham: Springer International Publishing), 555–572.

Nagel, E., and Newman, J. R. (2001). *Gödel's Proof*. New York, NY: NYU Press.

Nguyen, A., Yosinski, J., and Clune, J. (2015). "Deep neural networks are easily fooled: high confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 427–436.

Okasha, S. (2016). *Philosophy of Science: A Very Short Introduction*, 2nd Edn, Oxford: Oxford University Press.

Ord, T. (2020). *The Precipice*. New York, NY: Hachette Books.

Peirce, C. (1934). *Collected Papers, Vol 5.* Cambridge, MA: Harvard University Press.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford: Oxford University Press.

Pfeifer, R., and Bongard, J. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence*, Cambridge, MA: MIT Press.

Pfeifer, R., and Scheier, C. (2001). *Understanding Intelligence*, Cambridge, MA: The MIT Press.

Piaget, J. (1967). *Biologie et Connaissance*, Paris: Delachaux & Niestle.

Prokopenko, M. (2013). *Guided Self-Organization: Inception*, Vol. 9. Berlin: Springer Science & Business Media.

Ray, T. S. (1992). "Evolution and optimization of digital organisms," in *Scientific Excellence in Supercomputing: the 1990 IBM Contest Prize Papers*, eds K. R. Billingsley, H. U. Brown, and E. Derohanes (Atlanta, GA: Baldwin Press), 489–531.

Roitblat, H. (2020). *Algorithms Are Not Enough: Creating General Artificial Intelligence*. Cambridge, MA: MIT Press.

Roli, A., and Kauffman, S. (2020). Emergence of organisms. *Entropy* 22, 1–12. doi: 10.3390/e22101163

Rosen, R. (1958a). A relational theory of biological systems. *Bull. Math. Biophys.* 20, 245–260. doi: 10.1007/BF02478302

Rosen, R. (1958b). The representation of biological systems from the standpoint of the theory of categories. *Bull. Math. Biophys.* 20, 317–341. doi: 10.1007/BF02477890

Rosen, R. (1959). A relational theory of biological systems II. *Bull. Math. Biophys.* 21, 109–128. doi: 10.1007/BF02476354

Rosen, R. (1972). "Some relational cell models: the metabolism-repair systems," in *Foundations of Mathematical Biology, Vol. II*, ed R. Rosen (New York, NY: Academic Press), 217–253.

Rosen, R. (1991). *Life Itself: A Comprehensive Inquiry Into the Nature, Origin, and Fabrication of Life*. New York, NY: Columbia University Press.

Rosen, R. (2012). *Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations*, 2nd Edn. New York, NY: Springer.

Russell, S., and Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*, 4th Global Edition. London: Pearson.

Sanjuán, M. (2021). Artificial intelligence, chaos, prediction and understanding in science. *Int. J. Bifurc. Chaos* 31:2150173. doi: 10.1142/S021812742150173X

Scharmer, O., and Senge, P. (2016). *Theory U: Leading From the Future as It Emerges*, 2nd Edn, Oakland, CA: Berrett-Koehler Publishers.

Schneier, B. (2021). "The coming AI hackers," in *International Symposium on Cyber Security Cryptography and Machine Learning* Berlin: Springer, 336–360.

Searle, J. R. (1980). Minds, brains, and programs. *Behav. Brain Sci.* 3, 417–424. doi: 10.1017/S0140525X00005756

Searle, J. R. (1992). *The Rediscovery of the Mind*. Cambridge, MA: Bradford Books.

Shanahan, M. (2015). *The Technological Singularity*. Cambridge, MA: MIT Press.

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature* 529, 484–489. doi: 10.1038/nature16961

Skewes, J. C., and Hooker, C. A. (2009). Bio-agency and the problem of action. *Biol. Philosophy* 24, 283–300. doi: 10.1007/s10539-008-9135-9

Standish, R. K. (2003). Open-ended artificial evolution. *Int. J. Comput. Intell. Appl.* 3, 167–175. doi: 10.1142/S1469026803000914

Taylor, A., Elliffe, D., Hunt, G., and Gray, R. (2010). Complex cognition and behavioural innovation in new caledonian crows. *Proc. R. Soc. B Biol. Sci.* 277, 2637–2643. doi: 10.1098/rspb.2010.0285

Uexküll von, J. (2010). *A Foray Into the Worlds of Animals and Humans: With a Theory of Meaning*. Minneapolis, MN: University of Minnesota Press.

Varela, F., Maturana, H., and Uribe, R. (1974). Autopoiesis: the organization of living systems, its characterization and a model. *Biosystems* 5, 187–196. doi: 10.1016/0303-2647(74)90031-8

Vinge, V. (1993). "The coming technological singularity: how to survive in the post-human era," in *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace, NASA Conference Publication CP-10129*, ed G. A. Landis (Cleveland, OH: NASA Lewis Research Center), 11–22.

Walsh, D. (2015). *Organisms, Agency, and Evolution*. Cambridge: Cambridge University Press.

Whitehead, A. N. (1929). *Process and Reality*. New York, NY: The Free Press.

Wimsatt, W. (2007). *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Boston, MA: Harvard University Press.

Yudkowsky, E. (2008). "Artificial intelligence as a positive and negative factor in global risk," in *Global Catastrophic Risks*, eds N. Bostrom, and M. M. Cirkovic (Oxford: Oxford University Press).

Zaman, L., Meyer, J. R., Devangam, S., Bryson, D. M., Lenski, R. E., and Ofria, C. (2014). Coevolution drives the emergence of complex traits and promotes evolvability. *PLoS Biol.* 12:e1002023. doi: 10.1371/journal.pbio.1002023

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

REVIEW
published: 04 May 2022
doi: 10.3389/fevo.2022.796413

From Analog to Digital Computing: Is *Homo sapiens*' Brain on Its Way to Become a Turing Machine?

Antoine Danchin[1]* and André A. Fenton[2,3]

[1] School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong, Hong Kong SAR, China, [2] Neurobiology of Cognition Laboratory, Center for Neural Science, New York University, New York, NY, United States, [3] Neuroscience Institute at the NYU Langone Medical Center, New York, NY, United States

The abstract basis of modern computation is the formal description of a finite state machine, the Universal Turing Machine, based on manipulation of integers and logic symbols. In this contribution to the discourse on the computer-brain analogy, we discuss the extent to which analog computing, as performed by the mammalian brain, is like and unlike the digital computing of Universal Turing Machines. We begin with ordinary reality being a permanent dialog between continuous and discontinuous worlds. So it is with computing, which can be analog or digital, and is often mixed. The theory behind computers is essentially digital, but efficient simulations of phenomena can be performed by analog devices; indeed, any physical calculation requires implementation in the physical world and is therefore analog to some extent, despite being based on abstract logic and arithmetic. The mammalian brain, comprised of neuronal networks, functions as an analog device and has given rise to artificial neural networks that are implemented as digital algorithms but function as analog models would. Analog constructs compute with the implementation of a variety of feedback and feedforward loops. In contrast, digital algorithms allow the implementation of recursive processes that enable them to generate unparalleled emergent properties. We briefly illustrate how the cortical organization of neurons can integrate signals and make predictions analogically. While we conclude that brains are not digital computers, we speculate on the recent implementation of human writing in the brain as a possible digital path that slowly evolves the brain into a genuine (slow) Turing machine.

Keywords: recursion, cortical layers, micro-columns, learning, memory, algorithm

OPEN ACCESS

Edited by:
Giorgio Matassi,
FRE 3498 Ecologie et Dynamique des Systèmes Anthropisés (EDYSAN), France

Reviewed by:
Andrea Roli,
University of Bologna, Italy
Marcin J. Schroeder,
Tohoku University, Japan

****Correspondence:***
Antoine Danchin
antoine.danchin@normalesup.org;
adanchin@hku.hk

Specialty section:
This article was submitted to Models in Ecology and Evolution, a section of the journal Frontiers in Ecology and Evolution

Received: *16 October 2021*
Accepted: *14 March 2022*
Published: *04 May 2022*

Citation:
Danchin A and Fenton AA (2022) From Analog to Digital Computing: Is Homo sapiens' Brain on Its Way to Become a Turing Machine? Front. Ecol. Evol. 10:796413. doi: 10.3389/fevo.2022.796413

INTRODUCTION

The present essay explores key similarities and differences in the process of computation by the brains of animals and by digital computing, by anchoring the exploration on the essential properties of a Universal Turning Machine, the abstract foundation of modern digital computing. In this context, we try to explicitly distance XVIIIth century mechanical automata from modern machines, understanding that when computation allows recursion, it changes the consequences of determinism. A mechanical device is usually both deterministic and predictable, while computation involving recursion is deterministic but not necessarily predictable. For example, while it is possible to design an algorithm that computes the decimal digits of π, the value of any finite sequence following the *n*th digit, cannot (yet) be computed, hence predicted, with *n* sufficiently large. This implies that the consequences of replacing feedback (a common principle in mechanics) with recursion (a much deeper process, using a program that calls itself) are not yet properly addressed because they do not belong to widely shared knowledge. It is remarkable that recursion, associated with appropriate energy management, *creates* information (Landauer, 1961; Hofstadter et al., 1979). How this happens has been illustrated by Douglas Hofstadter in his book *Gödel, Escher, Bach, An Eternal Golden Braid*, as what he named a "strange loop," illustrated by a painting in an art gallery representing a person contemplating the painting in that very gallery. This illustration shows how a completely open new world where paradoxes are the rule is emerging (Hofstadter, 2007).

However, this happens on condition that a material support is involved, introducing a certain level of analogical information even in electronic computers. This involvement of the basic currencies of Reality other than information (mass, energy, space and time) opens up computing to another universe. This has consequences very similar to the result of Gödel's demonstration that arithmetic is incomplete: nothing in the coded integers used in the demonstration can say, within the number system, that there is a contradiction that can never be solved. It is only by going outside the coded system (so as to be able to observe it) that one can see the incompleteness. The very fact that the outcome of the demonstration can only be understood outside the frame of its construction—namely in a world where judgments exist—introduced a certain level of analogical information into the picture. The meaning of Gödel's last sentence (*I cannot be proved*) is not valid within the framework of the axioms and definitions of Number Theory, but only when one looks at Number Theory from the outside. Recursion is possible even in a world where analogical computing dominates, and the structure of the brain, organized in cortical layers, and through feedforward and feedback loops, may well allow the development of this procedure. However, the introduction of language, and of writing in particular, could well allow the modern human brain to behave like a Turing Machine, thus explaining how *Homo sapiens* could generate demonstrations of the type of Gödel's incompleteness theorems (Hofstadter et al., 1979).

ANALOG AND DIGITAL COMPUTING

We use "digital" to describe information and computation involving explicit numerical representations and manipulations, no matter how the numbers are themselves represented. In contrast, "analog" as it refers to neuronal information representation and computation, means a biological or other physical process like an action potential, that has another (i.e., is analogous to some) representational meaning or manipulation. In order to know whether thinking of the "brain as a computer" is more than a metaphor, we need to agree on a description of computing. We generally assume that computing involves an abstract process, the manipulation of integers with the standard rules of arithmetic. In this context the number *three* lies in an abstract world, beyond the way it is denoted: *trois* in French, *drei* in German, τρια in Greek, 三 in Chinese, etc. It belongs to the abstract domain of "information." This conception is based on the assumption that information is a true physical currency of reality (Landauer, 1996), along with mass, energy, space and time, allowing us to work in the abstract domain of digital computing. We must recognize, however, that the very concept of information, although widely used as a word, is not usually considered by biologists as an explicit physical entity, although some synaptic physiologists may conjecture that neural information is embodied in synaptic "strengths." As a consequence, when it comes to describing the role of the cell or the brain in computation, we have to oscillate between deep abstraction and concrete physiology. With a little more insight, when we use electronic computers, we combine Number Theory (Rosen, 2011) with the rules of logic, in particular Boolean logic (Sikorski, 1969). The vast majority of computing approaches are simply based on manipulation of bits. Computer users create algorithms, by subsuming a binary frame of reference, usually referred to as "digital" as a consequence of the way we calculate in the decimal system. However, the construction of relevant digital processing units asks for the understanding of the consequences of recursion and therefore Number Theory. Indeed, the consequences of Gödel's theorems makes that it remains impossible to design a processor which would be "hacking-free." This requirement is visible in constructs such as those belonging to the class of Verifiable Integrated Processor for Enhanced Reliability [VIPER, (Brock and Hunt, 1991)]. This vision remains a very crude abstract view of what computing is. It is based entirely on a discontinuous, discrete conception of physical reality. In contrast, the material world behaves as if it were continuous. In biology, the biochemical networks used in synthetic biology constructs are based on processes that amplify, synchronize, integrate signals and store information in a continuous way. Even computers are not exempt from the constraints imposed by the material world. The computer you are using to read this text is a material machine, made of components that have mass and obey the laws of physics. For example, the rules of logic are implemented as thresholds operating on continuous parameters, and even the very definition of a threshold cannot be entirely digitized. For example, it displays an inherent variability due to thermal noise. The history of digital computing acknowledges that, in parallel with an authentic shift toward a digital world possibly beginning with the ENIAC in 1946, computing kept being developed with analog devices (Misa, 2007).

In that sense, computation can be seen as both analog and digital. The idea of analog computation is not new. It seems to have been present very early, even in the ancient Greek civilization. In Greek, ανάλογος means "proportionate" with the notion that the due proportions associated with solving a certain problem can be used to solve that same problem *via* its simulation, not necessarily requiring understanding. The efficiency of analog computing is strikingly illustrated by an extraordinary device built more than 2,000 years ago, which seems to work like an analog computer to calculate a large number of properties of meteors in the sky, planets and stars, the *Antikythera Mechanism* (Freeth et al., 2021).

Today, the central processing unit that runs computers manipulates electrons, not cogs, in an organized way. It does not directly manipulate logical bits. How does the modern computer, which is constructed with components that have mass and space, perform its digital calculations? This may be a key question in our quest for the interaction between the analog world (that of matter with mass) and the digital world (that of the abstract genomic sequences manipulated by bioinformaticians for example) when we want to understand how the brain works. The question is indeed at the heart of what life is all about. How do we articulate the analog/digital interaction? We can recognize at least two different physical processes taking place in the electronic circuits of a digital computer. Continuous signals are transformed into digital (in fact usually Boolean) computation by: (1) exploiting the non-linearity of the circuits (transistors are either off or saturated, capacitors are either empty or fully charged). This entails that changes are rapid and large when compared to thermal fluctuations and means that we introduce thresholds such that a digital coding becomes reasonably robust. (2) Error correction mechanisms, such as redundancy, to overcome possible errors due to thermal electronic fluctuations. Essentially, this is the result of combining careful design of the basic electronic and physical phenomena with a coarse shaping (in a sense, a "clustering") and "constraining process" of the physical observables.

A frightening example illustrates the dichotomy between analog and digital information, not in the brain, but when we see cells as computers making computers, with their genetic program both analog, when it has to be accommodated within the cell's cytoplasm, and digital when it is interpreted as an algorithm for the survival of a cell and construction of its progeny. Viruses illustrate this dichotomy. Smallpox is a lethal virus. The sequence of its genome (digital information) is available, and can be exchanged *via* the Internet without direct action on the analog setup of living organisms, hence harmless. However, synthetic biology techniques (gene synthesis) allow this digital information to "transmute" into the analog information of the chemistry of nucleotides, regenerating an active virus. This coupling makes the digital virus deadly (Danchin, 2002). The difference between the textual information of the virus sequence and the final information of the finished material virus illustrates the complementarity between analog and digital computation. The analog implementation is extremely powerful, and we will have to remember this observation when exploring the way the brain appears to compute.

Physiological experiments meant to illustrate the first steps of computing in living systems are based on discrete digital designs but implemented in continuous properties of matter, such as concentration of ingredients. Typically, an early synthetic implementation of computing in cells, the toggle switch, consisted of the design of a pair of coding sequences of two genes, *lacI* and *tetR*, with relevant regulatory signals where the product of each gene inhibits the expression of the other (Gardner et al., 2000). This is possibly the simplest circuit capable of performing a calculation in a cell, in this case the ability to store one bit of information. Other digital circuits have been built since, and recent work has highlighted the level of complexity achieved by digital biological circuits, where metabolic constraints blur the picture. A deeper understanding of what happens in the cells where this construct has been implemented shows that their behavior is not fully digital (Soma et al., 2021). Cells, however, can still be seen as computers making computers and this performance can be described in a digital way (Danchin, 2009a). When we discuss the algorithmic view of the cell, we implicitly assume a digital view of reality. This is how we can introduce information, through "bits," i.e., entities that can have two states, 0 and 1. This is very similar to the way physics describes states as specific energy levels, for example (and this can be seen when an atom is illuminated, in the form of optically detected lines, which allows researchers to characterize the nature of that particular atom). However, because this oversimplified vision overlooks the analog dimension of computation, it omits taking into account material processes, such as aging for example, which requires specific maintenance steps involving specific functions that are rarely considered in digital machines [see (Danchin, 2015) and note that, in computers, processors also do age indeed, with important consequences on the computing speed and possibly accuracy (Gabbay and Mendelson, 2021)].

Unlike digital circuits, where a species has only two states, analog circuits represent ranges of values using continuous ranges of concentrations. In cases where energy, resources and molecular components are limited, analog circuits can allow more complex calculations than digital circuits. Using a relatively small set of components, Daniel et al. (2013) designed in cells a synthetic circuit that performs analog computations. This matches well with the common vision of synthetic biology, that of a cell factory which allows the expression of complex programs that can answer many metabolic engineering questions, as well as the development of computational capabilities. The take-home message of this brief discussion on the difference between analog and digital computing is that, in order to calculate, it is not necessary to do so numerically. Nature, in fact, appears as a dialog between continuous and discontinuous worlds. Some reject the idea of discontinuity. Forgetting about Number Theory, the French mathematician René Thom emphasized the continuous nature of the Universe and rejected the discontinuous view proposed by molecular biology, such as the way in which the genetic program is written as a sequence of nucleotides, acting as letters in a linear text written with a four-letter alphabet. He insisted on continuity even in the evolution of language (Thom et al., 1990), a point of view that we will discuss at the end of this essay. In summary, there are many facets of natural computing that we need to be aware of if we are to explore the brain computing metaphor (Kari and Rozenberg, 2008). We have seen how the *Antikythera Mechanism* was an early attempt using analog computing, and this line of engineering has been pursued over centuries. For example in 1836 a way was proposed to solve differential equations using a thread wrapped around a cylinder (Coriolis, 1836), and more recently using analog computers (Hartree, 1940; Little and Soudack, 1965; Barrios et al., 2019). Finally, after simulations of the behavior of the neuronal networks hardware was created that implemented analog computing into microprocessors (Wijekoon and Dudek, 2012; Martel et al., 2020).

VARIATIONS ON THE CONCEPT OF THE TURING MACHINE

Our digital computers are built according to an abstract vision, the Turing Machine (TM), elaborated by Alan Turing in the 1930s and developed in descriptions intended to make it more concrete by Turing and John von Neumann after the Second World War. The "machine" is often seen as a device reading a tape-like medium triggering specific behaviors, as we see them performed by computers. This view is rather superficial. It does not capture the key properties of the TM that created a general model of logic and computation, including the identification of impossibilities (Copeland, 2020). The machine is an abstract entity and, as in all other cases where we consider information as a genuine currency of physics, its implementation in objects with mass will create a considerable number of idiosyncratic constraints that can only be solved by what are sometimes called "kludges" in hardware machines, i.e., clumsy but critical solutions to a specific problem[1] [(Danchin, 2021b)]. This necessary overlap between information and (massive/spatial) matter creates the immense diversity of life, explaining why we witness so many "anecdotes" that interfere with our efforts to identify basic principles of life. Examples range from various solutions to the question opened by the presence of proline in the translation machinery, because proline is not an amino acid (Hummels and Kearns, 2020), to the need for a specific protease that cleaves off the first nine residues of the ribosomal protein L27 in the bacterial clade Firmicutes (Danchin and Fang, 2016), to macroscopic extraordinary display of color and behavior in birds of paradise (Wilts et al., 2014). It is important to remember that the way in which machine states could be concretely implemented, which marks the analog world, had no impact on the way the TM was used to contribute to the mathematical field of Number Theory (Turing, 1937). The "innards" of the machine were not taken into consideration.

The TM is a finite-state machine. In manipulating an abstract tape, it performs the following operations (the operations performed by computers are conceptually the same, although they appear to the general observer to be performed in a very detailed and therefore less comprehensible manner):

- Changing a symbol in a *finite* number of places, after reading the symbols found there (note that changing more than one symbol at a time can be reduced to a finite number of successive basic changes).
- Changing from the point which is being read to other points, at a given maximum distance away in the message.
- Changing the state of the machine.

All this can be summarized as specified by a series of quintuples, which each have one of the three possible following forms:

pαβLq or pαβRq or pαβNq

where a quintuple means that the machine is in configuration p, where symbol α is read, and is replaced by β to enter into configuration q, while displacing the reading toward the (L)eft, the (R)ight, or staying at the same (N)eutral place.

Several points need to be made here. The machine does not just read, it reads and *writes*. The machine can move forward, backward, and jump from one place to another. Thus, despite the impression that it uses a linear strip marked by sequences of symbols, its behavior is considerably more diverse. This is important when considering genomes as hardware implementations of a TM tape. For example, we tend to think of the processes of transcription, translation and replication as unidirectional, when in fact they are designed to be able to backtrack and change the building blocks that they had implemented in the forward steps, a process that is essential to their activity. More complex processes such as splicing and *trans*-splicing are also compatible with the TM metaphor. Furthermore, the above description corresponds to the Universal Turing Machine (UTM), which Turing showed to be equivalent to any construction using a finite multiplicity of tapes in parallel. A highly parallel machine can be imitated by a single-tape machine, which is of course considerably slower, but with exactly the same properties in terms of computational performance.

An essential point of the machine is that it is a finite-state machine. Despite its importance, this point is often overlooked, and it is here that the analogy between the cell or brain and a TM needs to be critically explored. What are and where are the states of the cell- (resp. brain-) machine located? Allosteric proteins have well-defined states, usually an active and an inactive state, and synapses are turned on or off, depending on the presence of effective neurotransmission. Their activity depends on the state of specific post-synaptic receptors, and often allosteric proteins (Changeux, 2013). More complex views may also take space into account, with the state of a protein or a complex defined by their presence at specific locations such as mid-cell or at the cell's poles for bacteria, or particular dendritic compartments of neurons (Hsieh et al., 2021). How do they evolve over time as the cell (resp. brain) "computes"? Of course, the same question can be asked of the hardware that makes up a computer, but in this case this is generally a key function of its memory parts, with specific addressing functions. We must also try to identify the vehicles that carry the information. In standard electronics, this role is played by electrons (with a specific role for the electric potential, which can move extremely quickly with an effect at long distance, whereas the physical movement of individual electrons is always slow). In optoelectronics, photons are used as information carriers, rather than electrons. What about cells? One could assume, in fact, that in most cases the information carriers are protons, which travel mainly on water molecules (forming hydrogen bonds) and on the surface of macromolecules and metabolites, also in forming hydrogen bonds (Danchin, 2021a). Part of the difficulty we have in visualizing what is going on in the cell is that we do not know well how water is organized, particularly around macromolecules, and how this might provide a series of hydrogen-bond channels carrying information from one place to another.

At this point the question becomes: how are the states of the cell fixed locally, i.e., what type of memory is retained, for how long and with which consequences? When we come to the

[1] https://www.theatlantic.com/technology/archive/2016/09/the-appropriately-complicated-etymology-of-kluge/499433/

way the brain computes, we will have to answer all these same questions. In the case of neurons, the carriers of information are primarily transmembrane potentials involving ionic currents that are initiated from dendrites, accumulate at the soma of neurons and after initiating action potentials propagate along the axons. But this is only part of the story: at synapses (except for electrical synapses), information is carried forward by specific neurotransmitters that trigger both ionotropic and metatropic neurotransmission, which introduces a strong coupling between information transfers and cellular metabolism (Chen and Lui, 2021), a feature that may be difficult to reconcile with the actual functioning of computers. It should be noted, however, that this organization creates *de facto* a range of relevant time frames that are considerably slower than the movements and state changes of the entities subject to thermal noise. This allows even short-term memories to be available for creation and recall without too much interference from temperature (recall that thermal vibrations typically occur in the femto-/pico-second range, while diffusion of a neurotransmitter occurs in the micro-millisecond range).

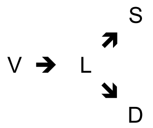
To transpose the TM concept into biology, further developments are needed in physics (what is information, how to represent it, etc.) and perhaps in mathematics (is there a need for mathematical developments other than number theory and logic, which are the basis of the formal description of TM and its parallel equivalents), in order to be able to embody it explicitly in soft matter. It is worth noting that, despite progress, there has not been much recent developments in Information Theories since the time of *Elements of Information Theory* (Cover and Thomas, 2006) and *Decoding Reality: The Universe as Quantum Information* (Vedral, 2012). The main problem is simple: in Turing's description, nothing is said about the machine, which is purely abstract, whereas it needs to be given some "flesh," with management of space, mass, time and energy.

Indeed, this tells us that there is a huge conceptual opportunity for recording the informational state of the cell as a TM (not only the transsynaptic cell membrane, cytoplasm, etc., but also the conformation of the chromosome, for example), which is much larger than the information carried by the genetic program as described as a sequence of abstract nucleotides (Danchin, 2012). Therefore, the transcription/translation machinery of the cell, as the concrete implementation of the mechanical part of the TM, the one that decides to move the program forward or backward, to read and write it, has enough opportunity to store and modify its states (its information). This is probably where the information retained by natural selection operates when cells multiply (and no longer just survive). Living systems can therefore act as information traps, storing for a time some of the most common states of the environment. Indeed, one might expect that what is involved in the machine's "decision" to progress (explore and produce offspring) is only a tiny subset of its information. From this point of view, natural selection seems to have a gigantic field of possibilities. We shall see that the problem is even more difficult to solve if we consider the brain. However, among the many features that characterize the TM, including the fact that it is a finite-state machine, it seems essential for systems that would ask to be recognized as a TM, to present distinct physical entities between the data/program set, and the machine that will interpret it into actions that modify the states of the machine.

BESIDES ENZYMES AND TEMPLATES: LEARNING AND MEMORY IN THE BRAIN

In 1949, Donald Hebb proposed that changes in the effective strength of synapses could explain associative learning (or conditioning, the process by which two unrelated elements become connected in the brain when one predicts the other). The idea was that the strength of a synapse could increase when the use of that synapse contributed to the generation of action potentials in the postsynaptic neuron (Hebb, 2002). With the intention of representing by an adequate formalism the learning phenomenon in the vertebrate central nervous system, based on Hebb's postulate, we have developed a theory of learning in the developing brain. This theory, implemented according to the axiomatic method, is placed within the general theory of systems where the nervous system is represented by a particular automaton. It is based on the idea of selective stabilization of synapses, depending on their activity (Changeux et al., 1973), phenomena that have been validated (Bliss and Collingridge, 1993; Bear and Malenka, 1994). The key to this vision (CCD model) is an epistemological premise: we seek to account for the properties of neural systems by means of a selective theory [in contrast with instructive theories, see discussion in Darden and Cain (1989)].

Based on families of experimental observations, this work restricted the study of memory and learning to neural networks (thus neglecting neuroglia and other features of the nervous system) and more specifically to the connections between neurons, the synapses. It proposed that, in addition to the now classic properties of networks traversed by impulses as found in computers (with the numerical logic rules that this imposes, as well as the feedback and feedforward loops), there is an original characteristic of neural networks, namely the possibility of a qualitative (and not only quantitative) evolution of synapses according to their activity. For simplicity, the model postulated that a synapse evolves, changing its state, in the graph:



where it passes during growth from a virtual state (V) to an unstable, labile state (L), then, depending on its local activity and the general activity of the posterior neuron, can either regress and disconnect (D) or stabilize in an active form (S).

With these very general premises, it was shown that a neural network is able to acquire the stable associative ability to recognize the form of afferent signals after a finite time. This memory and learning capacity comes from a transformation of the connectivity during the operation of the network. Thus, a temporal pattern is stored in the nervous network as a geometric spatial form. Besides quantitative involvement of synapse efficiency, the main originality of the approach lies in the fact that learning comes from the *loss* of connectivity. Learning carves a figure in the brain tissue that is memorized as a neural network. Moreover, even if the genetic constraints necessary to code for the implementation rules of this learning were very small, the system would nevertheless lead to the storage of an immense amount of information: each memorized event corresponds to a particular path traveling among the 10^{15} synapses of the network, so that the number of possible paths (and thus of memorized events) is combinatorially infinite. The only limit to our ability to learn—but it is a terribly constraining limit—is the slow access to our brain by our sensory organs, as well as the slow speed of the brain activity (as compared to that of modern computers, for example).

The CCD model explored the role of selective stabilization in learning and memory in the nervous system. This exploration preceded the fashion for neural networks, but with a twist rarely highlighted: living brain synapses evolved in such a way that they could regress and irreversibly disconnect from their downstream dendrite, making memorization irreversible at least for a time. In contrast, in most artificial neural networks, the state of the synapses is actually a quantitative feature that can revert to the initial values if the training set is noisy [see for example the initial model of the Perceptron (Lehtiö and Kohonen, 1978)]. As quantity is favored over quality, the latter has an important consequence: the outcome of the learning process is considerably sensitive to the length of the training period. The positive learning outcome first increases in parallel with the training period, then stabilizes and then gradually decreases if the training continues.

The role of neural networks has been and still is the subject of a considerable amount of work. An important sequel was the idea of *Neural Darwinism* proposed by Gerald Edelman (1987). The central idea of this work is that the nervous system of each individual functions as a selective system composed of groups or neurons evolving under selective pressure as selection operates in the generation of the immune response and in the evolution of species. By providing a fundamental neural basis for categorizing things, the aim of this hypothesis was to unify perception (network inputs), action (network outputs) and learning. The theory also revised our view of memory as a dynamic process of re-categorization rather than a replicative storage of attributes. This has profound implications for the interpretation of various psychological states, from attention to dreams, and of course, for the brain's computational capacity. Many other models of the links between memory, learning and computation have been proposed in recent decades. Most of them are based on neural networks, showing individual and collective behaviors with interesting properties that are not discussed here [see some eclectic examples in a vast literature (Dehaene and Changeux, 2000; Miller and Cohen, 2001; Mehta, 2015; Chaudhuri and Fiete, 2016; Mashour et al., 2020; Tsuda et al., 2020)].

To return to our question, can the brain be described as a computer, the basic idea behind these developments is that groups of neurons can allow the emergence of global behaviors while respecting the local organization of specific brain architectures. However, in general, this is mere conjecture, as there is no explicit demonstration of the behavior of the postulated structures. Nevertheless, this has triggered the emergence of a multitude of artificial neural networks (ANNs) that have developed metaphorically, independently of our knowledge of the brain. It is therefore interesting to see briefly how computation with neural networks has been implemented, which may now lead to a re-evaluation of their renewed link to brain behavior.

NEURAL NETWORKS

Simulation vs. Understanding

The work just mentioned is all centered on the interconnections of neurons, with the key view that neural networks are the objects that we should prioritize, before understanding the cell biology of neurons. In an ANN, a neural program is given which takes into account the "genetic" data (the "genetic envelope") of the phenomenon, the geometric data (essentially the maximum possible graph of all connections compatible with the genetic program, as well as, in some cases, the length of the axons in the form of propagation delays of the nerve impulse between one synapse and the next) and the operating data. Each neuron displays an integration function which, depending on the multi-message (afferent *via* the neuron dendrites), specifies the efferent message and the evolution function which, for each synapse, specifies its evolution toward a stable functional state which can be quantified. Again, in the CCD model a key property was that functioning under a genetically-programmed threshold led the synapse to evolve toward a degenerated non-functional state, thus disappearing as a connection. These processes depended on the afferent multi-message, as well as on a temporal law taking growth into account, i.e., the emergence of a new synapse in a functional state. During the operation of the system, a realization of the neural program is obtained at each time, which represents the effective anatomy of the network at that instant as well as its internal functioning.

Since it is quite difficult to understand the internal behavior of networks, especially when they consist of a large number of individual elements, neural networks have been studied mainly by modeling. In the absence of precise biological data, it has been necessary to propose hypotheses on the neural program data, especially regarding the function and structure of synapses. It has not yet been possible to create a detailed model of the synapse based on plausible physicochemical assumptions, so very approximate assumptions have been proposed for the integration and evolution functions of the neuron. The consequence is that, in general, the path followed is the development of ANNs that do not really mimic authentic neural networks. They are implemented as algorithms and then used with fast computers. It should be noted that ANNs can be trained to perform arithmetic operations with significant accuracy. Models of neural arithmetic logic units keep being continuously improved (Schlör et al., 2020). Whether such structures can be explicitly observed in authentic neural circuits remains to be seen. They often produce remarkably interesting results, but at a cost: it is not possible to understand how they achieved their performance. Many achievements made headlines, especially after AlphaGo beat European Go champion Fan Hui in 2015 and then Korean champion Lee Sedol in 2016. This demonstrated that deep learning techniques are extremely powerful. They continue to be developed by improving the structures and functioning of various networks (Silver et al., 2018; Czech et al., 2020). The use of these neural networks is currently limited to image or shape analysis or related diagnostic methods based on recognition of generally imperceptible patterns. As classic examples, these networks are used for making classes of objects, protein function prediction, protein-protein interaction prediction or *in silico* drug discovery and development (Muzio et al., 2021).

Unfortunately, successful predictions do not provide an explanation of the underlying phenomena, but only a phenomenological simulation of the process of interest, i.e., a process aimed at reproducing the observables we have chosen of a given phenomenon. These approaches, while extremely useful for diagnostic purposes, are unable to distinguish correlation from causation. To make the most of ANNs and use them as an aid to discovery, the result of their operation must be traceable in a causal chain. This restriction explains why legal regulators, in particular in the European Union, now require creators of AI-based models, often based on ANNs, to be able to demonstrate the internal causal chain of their successful models. This is understood as a way to associate prediction with understanding [[2] for an example of the way understanding can be visualized in an AI model, see for example Prifti et al. (2020)]. A major reason for the difficulty in tracing causality is the sheer size of networks required to perform simple tasks. For example, a simple visual image involves at least a million neurons in object-related cortex and about two hundred million neurons in the entire visual cortex (Levy et al., 2004). In this context, understanding causal relationships is often related to the ability of ANNs to generate systematic errors [see e.g., (Coavoux, 2021)], while error identification and correction is also important as it relates to intrinsic vulnerabilities against attacks, with the concomitant generation of spurious results (Comiter, 2019).

Neural Networks Organization: Cortical Layers

The fundamental organization of the cerebral cortical circuit of vertebrates remains poorly understood. In particular, it is not fully clear whether the considerable diversity of neuron types (Hobert, 2021) always form modular units that are repeated across the cortex in a way similar to what is observed in the cerebellum for example [Brain Initiative Cell Census Network [BICCN], 2021; Farini et al., 2021; Kim and Augustine, 2021]. The cortex of mammals has long been perceived as different from that of birds, in particular because in birds the folding of the cortical surface is particularly marked, but it now appears that the general organization in neuronal layers is quite similar in both phyla (Ball and Balthazart, 2021). This may be related to similar aptitudes in cognition/computation. There are so many models and conjectures about the role of the brain tissue organization that we had to make a choice for this essay. We will use the description/conceptualization proposed by Hawkins and Blakeslee in their book *On Intelligence. How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines* (Hawkins and Blakeslee, 2005) because it provides a compelling description of how the brain might work, notwithstanding the identification of new or alternative brain structures and functions over time. The title of the book comes from the idea that the cerebral cortex is composed of repeated micro-columns of microcircuits stacked side by side that cooperate to generate cognitive capacity. The book proposes that each of these columns has a good deal of innate capacity ("intelligence"), but only very partial information of the overall context. Yet, the cortical columns work together to reach a consensus about how the world works.

Hawkins and Blakeslee pictured the cortex: *as a sheet of cells the size of a dinner napkin, and thick as six business cards, where the connections between various regions give the whole thing a hierarchical structure*. An important feature in this description is its hierarchical organization, a feature identified as critical since the early work of Simon (1991). The cerebral cortex of mammals comprises six layers of specific neurons organized into columns. Layers are defined by the cell body (soma) of the neurons they contain (Shamir and Assaf, 2021). About two and a half millimeters thick, they are composed of repetitive units (Wagstyl et al., 2020). The strongest connections are vertical, from cells in the upper layers to those in the lower layers and back again. Layers seem to be divided into micro-columns, each about a millimeter in diameter, which function semi-independently, as we discuss below. The outermost layer of the neocortex, Layer I, is highly conserved across cortical areas and even species. It is the predominant input layer for top-down information, relayed by a rich and dense network of long-range projections that provide signals to the branches of the pyramidal cell tufts (Schuman et al., 2021). Layer II, is an immature neuron reservoir, important for the global plasticity of the brain connections. Within the view of the CCD model, it is an important place where synapses are expected to emerge from a virtual to a labile state. This layer contains small pyramidal neurons and numerous stellate neurons but seems dominated by neurons that remain immature even in adulthood, being a source of considerable plasticity (La Rosa et al., 2020). Pyramidal cells of different classes are predominant in layer III. In addition, multipolar, spindle, horizontal, and bipolar cells with vertically oriented intra-cortical axons are present in this layer. It also contains important inhibitory neurons and receives connections from adjacent and more distant columns while projecting to distant cortical areas. This layer has been explicitly implicated in learning and aging (Lin et al., 2020). Layers I-III are referred to as supragranular layers.

Layer IV is another site of cortical plasticity. It contains different types of stellate and pyramidal cells, and is the main target of thalamocortical afferents that project into distinct areas of the cortex, with, at the molecular level, specific involvement of phosphorylation regulatory cascades (Zhang et al., 2019). The major cell types in cortical layer V form a network structure combining excitatory and inhibitory neurons that form radial micro-columns specific to each cell type. Each micro-column functions as an information processing

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

unit, suggesting that parallel processing by massively repeated micro-columns underlies various cortical functions, such as sensory perception, motor control and language processing (Hawkins and Blakeslee, 2005; Hosoya, 2019). Interestingly, the micro-columns are organized in periodic hexagonal structures, which is consistent with the planar tiling of a layered organization (Danchin, 1998). Individual micro-columns are organized as modular synaptic circuits. Three-dimensional reconstructions of anatomical projections suggest that inputs of several combinations of thalamocortical projections and intra- and *trans*-columnar connections, specifically those from infragranular layers, could trigger active action potential bursts (Sakmann, 2017). Layer VI contains a few large inverted and upright pyramidal neurons, fusiform cells and a specific category, von Economo neurons, characterized by a large soma, spindle-like soma, with little dendritic arborization at both the basal and apical poles, suggesting a significant role of bottom up inputs (González-Acosta et al., 2018). This layer sends efferent fibers to the thalamus, establishing a reciprocal interconnection between the cortex and the thalamus. These connections are both excitatory and inhibitory and they are important for decision making (Mitchell, 2015).

Integrating Inputs and Outputs

The brain is connected to the various organs of the body. The sense organs provide it with information about the environment, while the internal organs allow it to monitor the states of the body, both in space and in time. This family of inputs is distributed in different areas of the brain, connected to cortical layers organized to integrate these inputs and allow them to drive specific outputs, in particular motor outputs (O'Leary et al., 2007). In this general structural signal processing, signals that reach a specific area of the brain connected to a given receptor organ pass through other areas, with feedback signals to connect to other sense organs. Locally, the integration structures of the brain are the micro-columns covering the six layers just described. The layered organization results in a limited number of neurons that integrate signals from other layers and parallel columns. In many cases signal integration may end up in a single cell, giving rise to the disputed concept of "grandmother cell," individual neurons that would memorize complex signals, such as the concept of one's grandmother or famous individuals like Halle Berry and Jennifer Aniston (Hawkins and Blakeslee, 2005; Quiroga et al., 2005, 2008; Bowers et al., 2019). To place this controversy in perspective, note that even complex brains can assign vital functions to individual neurons. For example, the deletion of a single neuron in a vertebrate brain abolishes essential behavior forever: the giant Mauthner cell, the largest known neuron in the vertebrate brain, is essential for rapid escape, so its loss means that rapid escape is also lost forever (Hecker et al., 2020).

The details of the integration of the input signals have been explored by Hawkins and Blakeslee, who provided an overview with a plausible scenario. The idea is that the individual columns are trained by experience *via* selective stabilization to represent and memorize particular families of environmental features. This implies that they encode invariant properties that can be used as a substrate to store and make invariant "predictions" (i.e., anticipation of future behavior) related to those particular features, from top to bottom (layer I to layer VI). Now, when the brain receives a particular input that matches one of these predictions, rather than triggering the activity of all the columns that represent similar features, it can be prompted to make an explicit individual prediction *via* a feedforward input that feeds the columns from the bottom up, consistent with the anatomy of cortical layer VI (**Figure 1**). This view is illustrated by Hawkins with the following image that shows how convolution of top/down and bottom/up inputs may result in a meaningful output. Imagine two sheets of paper with many small holes in them. The holes on one paper represent the columns that have active layer II or layer III cells, marking invariant predictions. The holes on the other paper represent columns with partial inputs from below. If you place one sheet of paper on top of the other, some holes will line up, others will not. The holes that line up represent the columns that should be active in making a specific prediction. This mechanism not only allows specific predictions to be made, but also resolves ambiguities in sensory input. This bottom/up top/down matching mechanism allows the brain to decide between two or more interpretations and to anticipate events that it has never witnessed before. Further developments of the modular organization of the animal brain may have developed with the emergence of *Homo sapiens*, resulting in specific amplification of new connection modules (Changeux et al., 2021). This important behavior results from an organization that combines the columns into layers with overlapping lateral connections, a feature we explore later.

Synchronization

Finally, a critical feature for computation is the need for synchronization of processes. Understanding how cortical activity generates sensory perceptions requires a detailed dissection of the function of time in cortical layers (Adesnik and Naka, 2018). This is the case, for example, of the eye saccade movement that controls vision, allowing proper positioning of the retina to keep proper focus while the eye moves (Girard and Berthoz, 2005). Synchronization is important not only with single computing units but especially important when computing is developed in parallel. For this reason it seems relevant, before understanding whether brain computing can become digital, to identify at least some families of synchronization processes. Many regular waves, spanning a wide time frame, have been identified in the brain, witnessing large scale synchronization processes, particularly important for information processing in virtually all domains including sensation, memory, movement and language (Buzsáki, 2010; Meyer, 2018). Time keeping can be achieved for example *via* coupling two autonomous dynamic systems (Pinto et al., 2019). Recent work follows older work where small populations with a feedback loop were shown to mimic the behavior of authentic neural networks (Zetterberg et al., 1978). In addition, the need to make use of the states that have been stored requires a scanning process that is essential to enable functions such as memory recall that is distinct from encoding the information from experience (Dvorak et al., 2018).

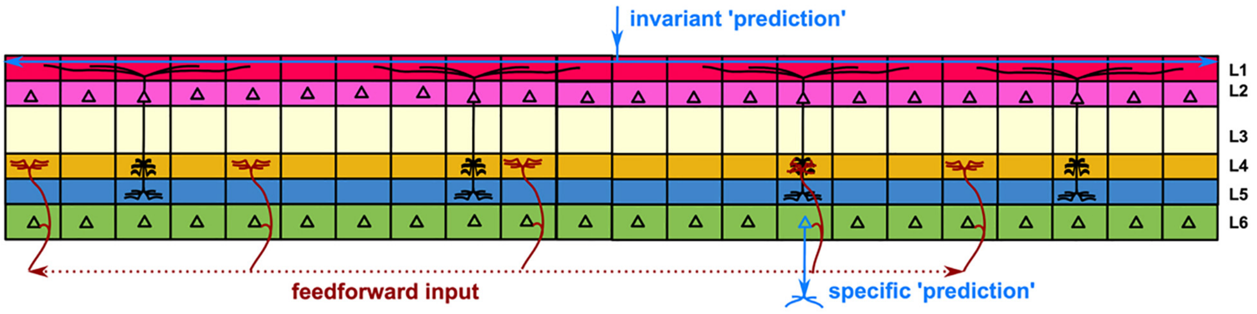


FIGURE 1 | Redrawn from Hawkins and Blakeslee (2005). Formation of a specific "prediction" in the cortex. The cortex is represented by six layers connecting micro-columns. Generic "predictions" result from memories entered in the columns of the supragranular layers (see text) and triggered by interaction with the environment in a top-down manner. To obtain a specific "prediction" that will result in a specific output, a family of anticipatory feedforward signals is input from the lower layers in a bottom-up manner. The convolution of the descending and ascending signals produces the specific output.

Indeed, the neural oscillations observed in local field potentials that result from spatially and temporally synchronized excitatory and inhibitory synaptic currents (Buzsáki et al., 2012) provide powerful network mechanisms to segregate and discretize neural computations operating within a hierarchy of time scales such as theta (140 ms) cycles, within which (30 ms slow gamma and (14 ms fast gamma oscillations are nested and theta phase organized. This temporal organization is intrinsic, arising from the biophysical properties of the transmembrane currents through ion-conducting channel proteins. The information processing modes within and between cortical processing modules that these oscillations enable are themselves controlled by top-down synchronous inputs such as medial entorhinal cortex-originating dentate spike events (Schomburg et al., 2014; Dvorak et al., 2021). By hierarchically synchronizing synaptic activations, the intrinsic biophysics of neural transmission accomplishes a remarkable form of digitization. Continuous inputs at the level of individual neurons are converted into oscillation-delineated population synchronized activity with digital features of a syntax for discretized information processing (Buzsáki, 2010), disturbances of which result in mental dysfunction (Fenton, 2015).

AN AUTOMATIC SCANNING PROCESS, THE UNEXPECTED BENEFIT OF FUZZINESS

As described above, large parts of the animal brain are organized as an association of local micro-networks of similar structure, arranged along planar layers and micro-columns. It is therefore of interest to identify the basic units that might play a role in this organization. Phylogenetic analyses are important in trying to identify functions of neuronal structures that appear for the first time in a particular lineage. Typically, at the onset of the emergence of animals, a neuron was a kind of relay structure that couples a sensory process to a motor process. At the very beginning of the development of such structures during the evolution of multicellular organisms, the role of neurons was simply to couple sensing with the movement produced by distant organs. However, this simple process is bound to have a variety of undesirable consequences if it does not resolve its role within a well-defined spatial and temporal framework. This means that the sequence associating the presence of a signal to its physiologic or motor consequence must be delimited in time and space. A relevant design to ensure the quality of this process is to divert a small part of the output to inhibit the effect of the upstream input, in short, to achieve a homeostat (Cariani, 2009).

Homeostasis: The Negative Feedback Loop

In his *Neural Darwinism* Edelman developed the concept of "reentry," a key mechanism for the integration of brain functions (Edelman and Gally, 2013). This concept is based on the idea that a small part of the output signal of a network is diverted to the input region and fed back into the network with a time delay. This phenomenon belongs to the family of signals that ensure homeostasis. A central theme governing the functional design of biological networks is their ability to maintain stable function despite intrinsic variability, including noise. In neural networks, local heterogeneities progressively disrupt the emergence of network activity and lead to increasingly large perturbations in low frequency neural activity. Many network designs can mitigate this constraint. For example, targeted suppression of low-frequency perturbations could ameliorate heterogeneity-induced perturbations in network activity. The role of intrinsic resonance, a physiological mechanism for suppressing low-frequency activity, either by adding an additional high-pass filter or by incorporating a slow negative feedback loop, has been successfully explored in model neurons (Mittal and Narayanan, 2021).

The cerebellum, with its highly regular organization and single-fiber output from Purkinje cells, is a good example of repetitive networks. Mutual inhibition of granule cells, mediated by feedback inhibition from Golgi cells—much less numerous than their granule counterparts—prevents simultaneous activation. Granule cells differentiate by their priming threshold, resulting in bursts of spikes in a "winner take all" sequential pattern (Bratby et al., 2017). Taken together, the local implementation of networks with embedded feedback loops as a strong output used to re-enter relevant cortical large networks resulted in a pattern that was proposed to explain the origins of consciousness and its scanning properties (Edelman et al., 2011), and further extended into the Global Neuronal Workspace hypothesis that attempts to account for key scientific observations regarding the basic mechanisms of conscious processing in the human brain (Mashour et al., 2020). These views, where inhibition is crucial, are strongly supported by the considerable importance of circuits comprising inhibitory neurons. Inhibition in the cortical areas is implemented by GABAergic neurons, which comprise about 20–30% of all cortical neurons. Witnessing the importance of this negative function, this proportion is conserved across mammalian species and during the lifespan of an animal (Sahara et al., 2012).

Finally, the role of inhibition, which typically occurs locally but is typically triggered by inputs from distant areas of the brain, is particularly important for the discrimination of classes of processes. When neural network excitatory inputs are both mutually excitatory and also recruit inhibition globally, the motif generates winner-take-all dynamics such that the strongest and earliest neural inputs will dominate and suppress weaker and later inputs, which in turn causes further enhancement of the dominant inputs. The net result is not merely a signal-to-noise enhancement of the dominant activity, but a network selection and discretization of what would be otherwise continuously variable activity. This motif is learned, improves with experience and in the entorhinal cortex-hippocampal circuit is responsible for learning to learn (Chung et al., 2021). The study of child brain development shows that there is a progressive overlap of organized responses to specific inputs (in the way objects and then numbers are identified) with other types of input from, for example, visual areas. The consequence of this overlap is that "intuitive" conceptions, resulting from prior anchoring in a particular environment, are barriers to conceptual learning. This implies that the inhibition of these inputs is important to allow the development of rationality (Brault Foisy et al., 2021). Interestingly, this duality between intuition and rational reasoning can be attributed to a difference between heuristics and algorithmic reasoning (Roell et al., 2019), a feature that may support the transition from a purely analog to a digital process.

Consequences of Imperfect Feedback: Endogenous Scanning of Brain Areas

The phenomenon of consciousness suggests that the brain generates an autonomous process that allows it to continuously scan the network, extracting information to promote action, at least during the waking period. It is therefore important to propose conjectures about how this process is generated. Many connection schemes using feedback or feedforward signaling are well suited to enable homeostasis, but there is a particularly simple one that seems to have interesting properties for producing scanning behavior. Suppose that at some level sensory inputs are split and thus follow parallel paths, only to be re-associated and, for example, because one of the cells in one path activates an inhibitory neuron, negatively controlling the downstream neuron, and then activates the neurons in the other path, subtracting one from the other only at the level of a specific class of cells, with a XOR-like local network (Kimura et al., 2011; Michiels van Kessenich et al., 2018). Measuring fine differences is a way to extract subtle information from the environment and make it relevant. Cells of the latter class are then assumed to return one of the duplicated sensory inputs or intermediate inputs corresponding to "modifications" of these inputs (see a metaphoric illustration in **Figure 2**). If the difference read by the cell integrating the commands from the two parallel pathways is very small, the feedback inhibition command will have no effect; if the difference is large, this command may cancel or reinforce one of the upstream pathways, so as to cancel the difference, thus resulting in homeostatic behavior. If the cell in the last layer remains activated, it tends on the one hand to produce an action *via* its connection to a motor center, and on the other hand to correct the influence of the input system that leads it to command the action.

Now, consider how these structures are built during brain development. Living matter, unlike standard inorganic matter, is soft matter. This intrinsic flexibility must be taken into account when considering the fine architecture of authentic neural networks. When we describe columns of cells organized into a hexagonal planar structure, it cannot consist of structures with precisely defined boundaries (Tecuatl et al., 2021). In another level of fuzziness, involving time, the effectiveness of individual synapses is not strictly defined, resulting in pervasive synaptic noise (Kohn, 1998). Moreover, the individuality of each synapse can only be programmed exceptionally as such: this would require at least one gene per synapse, and remember that there are at least 10^{15}. This implies that there is considerable variation in the temporal and spatial dependence of neuronal connectivity. Rather than being an obstacle, this weakness gives rise to a new strength: it is because neural networks cannot be programmed exactly to create precise homeostatic structures that they lead to the repetition of approximate structures that are quasi-homeostatic, creating interactions with their neighbors that can be used to implement emerging functions.

Indeed, such networks have the interesting property of being able to trigger an automatic network scanning process. When an input signal triggers a homeostatic response from one column meant to inactivate it after a time, it inevitably activates the response of adjacent columns to which it is connected because of inevitable variation in dendrites and axons connections. In turn, this initiates a homeostatic response evolved to silence them. In so doing they now activate adjacent columns, thus initiating a local scan of the memorized information stored in those columns, progressing by contiguity as a wave. In line with the notion of reentry, why not propose this process at the origin of consciousness? This metaphoric vision developed into a dialog between biology and the formal properties of syntactic structures proposed by Noam Chomsky [see Danchin/Marshall exchange in Modgil and Modgil (1987)]. Of course, an infinite number of variations on this theme, playing on the differences between nearly identical signals can act as a scanning process that will recall memories *via* the sequential activation followed by the inactivation of parallel structures. Since this process is

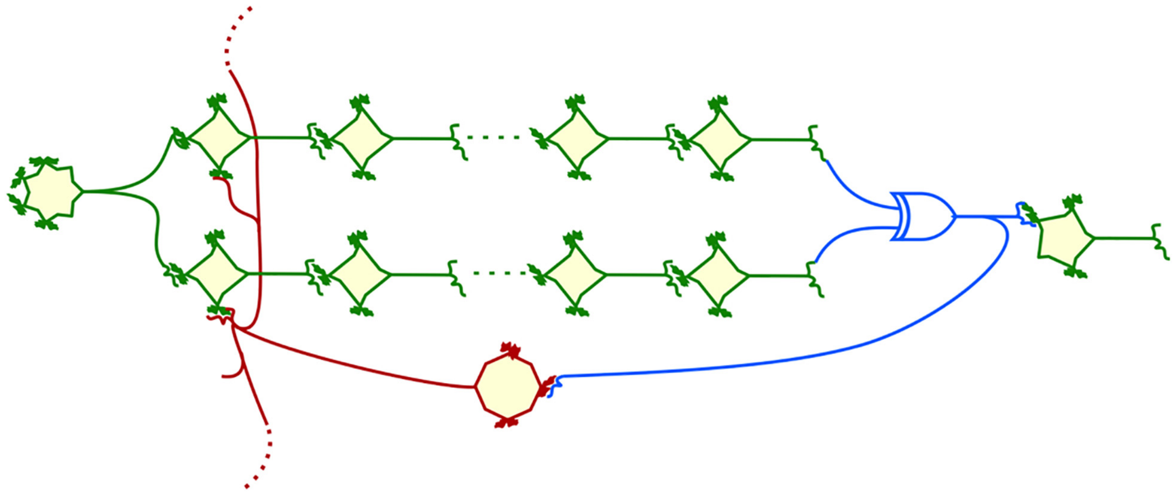


FIGURE 2 | A representation of a unit cell for a quasi-homeostat. An input cell is connected to an array of duplicated column cells that inputs in a microcircuit behaving as an XOR logic gate (blue) subtracting signals from the two parallel columns. A fraction of the network output is diverted to activate an inhibitory neuron (red) that feedbacks to the origin of the duplicated columns. Since the corresponding connections cannot be coded individually, they will also connect to adjacent columns and trigger their activity, initiating a scanning process.

spatially constrained by contiguity, it will give the recall of memories a spatial component, such as we all experience when we have to retrace our steps to find a memory that has just escaped our attention.

THE COMPUTING BRAIN

A TM must separate the machine and the data/program physically, noting that the data/program entity cannot be split into specific entities but belongs to a single category, that is processed by the machine to modify its state. Where does the brain fit in this context? Can we distinguish between a set of data/programs and the state machine that manages it? An interesting observation from an interview by the *Edge Magazine* with Freeman Dyson in 2001 gives us a hand in broadening our discussion[3]. Freeman Dyson, as usual, is an extraordinary mind: *The two ways of processing information are analog and digital. [.]. We define analog-life as life that processes information in analog form, digital-life as life that processes information in digital form. To visualize digital-life, think of a transhuman inhabiting a computer. To visualize analog-life, think of a Black Cloud* [reference to the novel of Fred Hoyle (1957)]. *The next question that arises is, are we humans analog or digital? We don't yet know the answer to this question. The information in a human is mostly to be found in two places, in our genes and in our brains. The information in our genes is certainly digital, coded in the four-level alphabet of DNA. The information in our brains is still a great mystery. Nobody yet knows how the human memory works. It seems likely that memories are recorded in variations of the strengths of synapses connecting the billions of neurons in the brain with one another, but we do not know how the strengths of synapses are varied. It could well turn out that the processing of information in our brains is partly digital and partly analog. If we are partly analog, the down-loading of a human consciousness into a digital computer may involve a certain loss of our finer feelings and qualities.*

An important feature that can be added to the question posed by Dyson is that the brain, through its learning process, constructs a hierarchical tree structure and symbolic links from the data submitted to it (basically, it throws away most of the data, condenses it into another form at a higher level of abstraction to sort and order it). Capturing the involvement of elusive information is difficult (we still do not have a proper formalism to describe what it is). The most common approach to tie information to energy has been proposed by Rolf Landauer and Charles Bennett, with the understanding that during computation, creation of information is reversible (hence does not dissipate energy) while erasing memory to make the result of computation stand out against the background costs *kTln2* per bit of information (Landauer, 1961; Bennett, 1988b). It is well established that the brain consumes a considerable amount of energy, but the relationships with information processing have not been investigated in-depth.

The word "program," often used loosely to describe the concrete implementation of a TM, implies the anthropocentric requirement of a goal. However, a TM does not have an objective, it is "declarative," i.e., it functions as soon as a tape carrying a string of data is introduced into its read/write machinery. Understanding how it works is therefore better suited to the idea of data not program manipulation. The distinction between data and program opens a difficult scene in the concept of information. Data has no meaning in itself, whereas the program depends on the context (Danchin, 2009b). This distinction is evident in the cell where genetic information duplicated during the process of DNA replication starts as soon as a DNA double

[3] https://www.edge.org/conversation/freeman_dyson-is-life-analog-or-digital

helix meets a DNA polymerase machinery. This process does not see the biological significance of the encoded genes or other features of the DNA sequence. This has been well established with *Bacillus subtilis* cells transformed with a cyanobacterial genome that is faithfully replicated but not expressed, whereas it drives the synthesis of an offspring when present in its parent host (Watanabe et al., 2012). This distinction between Shannon-like information (meaningless) and information with "value" has been discussed for a long time under the name "semantic" information [(Bar-Hillel and Carnap, 1953; Deniz et al., 2019; Lundgren, 2019; Miłkowski, 2021), see also emphasis on the requirement for recursive modeling to account for information in the brain (Conant and Ashby, 1970)]. However, with the exception of the idea of logical depth, proposed by Bennett in 1988 (Bennett, 1988a), there is still no well-developed theory on the subject. We will restrict our discussion to the role of data in the TM.

The possibility of moving from one set of data to a smaller set, as illustrated in the functioning of cortical layers, is quite similar to the measure proposed by Bennett when he illustrated Landauer's principle by the process of arithmetic division. In this illustration, Bennett showed that this operation could be implemented in a reversible way, leaving the remainder of the division as its result (Bennett, 1988b). However, in order to bring out the division remainder, to make it visible, it is necessary to erase all the steps that led to the result: this is what costs energy. What is indeed important is the sorting that allows the relevant data to be isolated from the background. To carry out a sorting, a choice, it is necessary to carry out a measurement, as Herbert Simon pointed out in his decision theory (Simon, 1974). In living cells, this process is fairly easy to identify in the process of discriminating between classes of entities, for example young and old proteins. In this case, the question is how to verify that the cell is measuring something before "deciding" to degrade a protein. Many ways of achieving this discrimination can be proposed. In a cell, the cleaning process could simply be a prey-predator competition between proteins and peptidases. It could be the result of spatial arrangement with producing or moving proteins in a place where there are few degrading enzymes, or the fact that when the proteins are functioning, they form a block, an aggregate, that is difficult to attack. All these processes dissipate energy at steps that specifically involve information management (Boel et al., 2019). But what about neural networks?

There is no doubt that the brain manipulates information and computes. But where are its states stored and how are they managed? Discrimination processes can easily be identified in the way the brain tackles its environment, but where do we find specific energy-dependent processes underlying discrimination? Furthermore, there does not seem to be any data/program entity that can be exchanged between brains. *Homo sapiens* is perhaps an exception, when true language has been established. Animal communication may also make use of the same observation, but less obviously, and certainly not if we follow the Chomskian definition of language (Hauser et al., 2002). Sentences can be exchanged between different brains, in a way that alters the behavior of the machine that carries the brain: this is particularly visible with writing, which is the metaphor used by Turing, but it is certainly true as soon as writing is established, which makes writing the benchmark of humanity.

WRITING: TOWARD A TURING MACHINE?

We enter here a very speculative section of this essay, meant to help generate new visions of the brain's competence and performance. In fact, while von Neumann and others invented computers with mimicking the brain in mind, the brain does not appear to behave as a TM. **Table 1** compares the key features of a Turing Machine, a computer, and the human brain (**Table 1**). In case we accept that the brain could work as a digital computer, several features of the digital world should be highlighted as they should have prominent signatures. Among those that have unexpected but recognizable consequences, we find recursion (Danchin, 2009a) and this fits with the concept of reentry. An original feature of the TM is that it allows recursion. Recursion is built into numerical worlds when a routine executes a program that calls itself. A consequence of recursion, which was addressed by Hofstadter in his *Gödel Escher Bach* (Hofstadter, 1999), is that it produces inherently creative behavior (i.e., giving rise to something that has no precedent), a feature commonly observed in the role of the brain. This happens in cells, even before they multiply (especially when they repair their DNA, during the stationary phase). The digital life of the cell provides a recursive way to creatively explore its future. Creation is also a key feature of TMs, and brains. Consider the wiring diagram of the mammalian brain comprised as it is of parallel cortico-striatal-thalamo-cortical loops each specialized for motor, visual, motivational, or executive functions (Alexander et al., 1986; Seger, 2006). These canonical loops provide the essential circuitry for recursive information processing, as observed in the creation of complex abstract rules from simple sensory motor sequences (Miller and Buschman, 2007). An elegant study stimulated the mossy fiber component of the cerebellar circuity within the additional parallel cortico-cerebellar-thalamo-cortical loop to causally demonstrate recursion in the formation of a classically conditioned eyeblink response (Khilkevich et al., 2018). Indeed, such recursion-implementing circuity is widespread and characteristic of the mammalian brain (Alexander et al., 1986).

Among the characteristics necessary to identify a TM is the physical separation of the data/program from the machine that interprets it. The data/program entity is illustrated as a string of symbols, a purely digital representation. The metaphoric string of the Universal TM can be embodied into a variety of strings (parallelization is allowed). A brain, on the other hand, seems to work with a completely different approach. There is no separate program involved in its operation. It is simply "programmed" by the interconnections between its active components, neurons. The brain does not appear to fetch instructions or data from a memory located in a well-defined area, decode and interpret instructions etc. Neurons get input data from other neurons, operate upon these data and generate output data that are fed to receiving neurons. Memory is distributed all over the brain tissue.

TABLE 1 | Some specific features of a Turing Machine, a computer, and the human brain.

	Turing Machine	Computer	Brain
Separation data/machine	Yes	Yes	No in general Yes for grammatical language performance and numbering
Declaration	Yes	No (not yet)	Yes
Prescription	No	Yes	Yes in social organisms
Digital	Yes	Yes	Limited to numbering and language performance
Analog	No	Yes for the machine	Yes
Recursive	Yes	Yes	Yes
States	Finite	Many finite states	Unlimited, poorly defined
Organizational granularity	Two separate levels	Multilevel	Multilevel
Memory	Past state	RAM and ROM	Distributed, limited only by lifespan
Computation	Algorithmic	Algorithmic (heuristics can be implemented *via* algorithms)	Heuristic and algorithmic

This view is a bit oversimplified but it is enough to bring us to the following questions. Do we find entities that can be separated from the brain, extracted and reintroduced, in the way it works? In an animal brain, the question is to understand what might play the role of strings of symbols.

Separation has a considerable consequence: it requires some kind of communication and exchange, which could occur between parts of the brain. This feature was discussed in an interesting essay by Julian Jaynes, where he surmised that consciousness emerged from a dialog between coded sequences between the brain hemispheres [our ancestors had "voices" (Jaynes, 2000)]. While this vision is now obsolete, it points out how this could be the first step of a pre-TM where the brain exchanges strings of signals between hemispheres (contemporary views consider areas rather than hemispheres, with particular emphasis on inhibition) to generate novel information. In fact, such phenomena have been experimentally demonstrated by interhemispheric transfer and interhemispheric synthesis of lateralized engrams, studies that exploited the ability to reversibly silence one and then the other of the brain's bilateral structures (Nadel and Buresova, 1968; Fenton et al., 1995). This organization of the cortex is also used by animals to map future navigation goals (Basu et al., 2021). Alongside this evolution of information transfer within the brain, strings of symbols could be exchanged between brains, implying that the social brain is at the heart of what is needed for a brain to become a TM. When language comes into play (probably first through grammatically organized phonemes and then through writing), part of the brain may behave as a digital device, with important properties derived from the corresponding TM scenario. This is what is happening now, when you read this text: your brain behaves like a TM, and you can modify the text, exchange it *via* someone else's brain, etc. In fact, some of this may already be true in the ability to see. The images seen by the retina are in a way digitized *via* the very construction of the retina as layers of individual cells, that "pixelize" the image of the environment. It is not far-fetched to assume that the processing of the corresponding information by brain neural networks has retained some of the characteristics of this digitization.

Before the invention of writing, making reusable tools would also represent a primitive way of implementing a TM. *Homo sapiens* is one of the very few animals to do so. In birds, tools can be made, but tool reuse and tool exchange have rarely been observed. It may therefore be that the genus Homo began to build a TM-like brain, but that its actual implementation as an important feature only appeared with complex languages (i.e., with grammatical properties linked to a syntax of the type described by Noam Chomsky) and, most importantly, with the invention of writing. Looking again at the network layer organization of the cortex, we can see that there is a certain analogy between the simplest elements of syntactic structures and this neural structure. The afferent pathways (and not the cells), all constructed in the same way (but not identical), would represent the nominal syntagm (with all that the numerous variants of afferents, interferences, modifications can bring to meaning) and the cells integrating their output commands, the verbal syntagm (with all that this implies in terms of motor actions, including imaginary ones, since by construction the verbal syntagm acts on the nominal syntagm.) This is a gradual evolution, which will certainly undergo further stages in the future. Invention of language with its linear sequences of phonemes, when spoken, and letters when written, would mark, in *Homo sapiens* the transition moment when it behaved as a Turing Machine and separate human beings from other animals. One caveat, though. The emphasis here has been on one of the characteristics of the TM, namely the physical separation of the data from the machine, where data can be replaced by other data without changing the specific nature of the machine. However, there is a second essential characteristic of a TM: it is a finite state machine. It would be difficult to accept that the brain behaves like such a machine. Even its states are quite difficult to identify (although progress in identifying the functioning of various areas may provide some insight into the localized features of specific states). It would be necessary to validate the hypothesis discussed here, to couple the way writing is used with specific states. This remains quite futuristic.

IN GUISE OF CONCLUSION: THE BRAIN IS NOT A DIGITAL COMPUTER, BUT IT COULD EVOLVE TO BECOME ONE

This essay is not intended to review the vast amount of work exploring the computational capacity of the human brain. We have extracted from the literature leads that have allowed us to find new answers to the question: is the brain a digital computer? As a final clue, we could ask whether we can be infected by a computer virus. It may sound far-fetched, but yes, it is apparently possible, as we see with fake news or memes that spread *via* social media. Both are, however, linked to the language processing ability of the human brain. Social media manage information with an explicit separation of data and a machine. This observation is reminiscent of the way the mind/body problem is asked: there is no "ghost in the machine" (Ryle, 2009), but nobody would doubt that brain manages information in a very efficient way. However, this affirmation does not tell us whether this is made in an analog or digital way. Nevertheless, the consequences for pedagogy of the algorithmic vision are considerable. Learning to read by "gazing" at written words is a matter of the brain's as an analog device vision, while reading by breaking words down into syllables is a matter of digital vision. Forgetting the algorithmic nature of the corresponding processes must have deep consequences in terms of the organization of the brain, and possibly jeopardize long term cognition abilities. It has been proposed that cells may act as computers making computers, with an algorithmic description of the cell's behavior based on the way macromolecules are synthesized, with the key role of the genetic code with the algorithmic description of decoding (Danchin, 2009a). With the view that the human brain might be on its way to become a TM, Nature would have discovered twice the importance of coding and recursion, in the emergence of cell life with the discovery of the genetic code, first, and in the emergence of writing, quite recently.

AUTHOR CONTRIBUTIONS

AD created the content of the article and wrote the manuscript. Both authors contributed to the article and approved the submitted version.

FUNDING

AF is supported by NIH grants R01MH115304 and R01NS195472.

ACKNOWLEDGMENTS

The present reflection is the result of collective scientific work. It has been pursued over the years and discussed in the weekly seminar of Stanislas Noria. Part of it is available on the website https://www.normalesup.org/~adanchin. The special role of the i2Cell program of the Fondation Fourmentin-Guilbert is especially acknowledged, with the contributions of Éric Fourmentin, Damien Larivière, and Jean-Luc-Marion.

REFERENCES

Adesnik, H., and Naka, A. (2018). Cracking the function of layers in the sensory cortex. *Neuron* 100, 1028–1043. doi: 10.1016/j.neuron.2018.10.032

Alexander, G. E., DeLong, M. R., and Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu. Rev. Neurosci.* 9, 357–381. doi: 10.1146/annurev.ne.09.030186.002041

Ball, G. F., and Balthazart, J. (2021). Evolutionary neuroscience: are the brains of birds and mammals really so different? *Curr. Biol.* 31, R840–R842. doi: 10.1016/j.cub.2021.05.004

Bar-Hillel, Y., and Carnap, R. (1953). Semantic information. *Br. J. Philos. Sci.* 4, 147–157. doi: 10.1093/bjps/IV.14.147

Barrios, G. A., Retamal, J. C., Solano, E., and Sanz, M. (2019). Analog simulator of integro-differential equations with classical memristors. *Sci. Rep.* 9:12928. doi: 10.1038/s41598-019-49204-y

Basu, R., Gebauer, R., Herfurth, T., Kolb, S., Golipour, Z., Tchumatchenko, T., et al. (2021). The orbitofrontal cortex maps future navigational goals. *Nature* 599, 449–452. doi: 10.1038/s41586-021-04042-9

Bear, M. F., and Malenka, R. C. (1994). Synaptic plasticity: LTP and LTD. *Curr. Opin. Neurobiol.* 4, 389–399. doi: 10.1016/0959-4388(94)90101-5

Bennett, C. H. (1988b). Notes on the history of reversible computation. *IBM J. Res. Dev.* 44, 270–277. doi: 10.1147/rd.441.0270

Bennett, C. H. (1988a). "Logical depth and physical complexity," in *The Universal Turing Machine, A Half Century Survey*, ed. R. Herken (Oxford: Oxford University Press), 227–257.

Bliss, T. V. P., and Collingridge, G. L. (1993). A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* 361, 31–39. doi: 10.1038/361031a0

Boel, G., Danot, O., de Lorenzo, V., and Danchin, A. (2019). Omnipresent Maxwell's demons orchestrate information management in living cells. *Microb. Biotechnol.* 12, 210–242. doi: 10.1111/1751-7915.13378

Bowers, J. S., Martin, N. D., and Gale, E. M. (2019). Researchers keep rejecting grandmother cells after running the wrong experiments: the issue is how familiar stimuli are identified. *Bioessays* 41:e1800248. doi: 10.1002/bies.201800248

Brain Initiative Cell Census Network [BICCN] (2021). A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* 598, 86–102. doi: 10.1038/s41586-021-03950-0

Bratby, P., Sneyd, J., and Montgomery, J. (2017). Sequential pattern formation in the cerebellar granular layer. *Cerebellum* 16, 438–449. doi: 10.1007/s12311-016-0820-y

Brault Foisy, L.-M., Ahr, E., Blanchette Sarrasin, J., Potvin, P., Houdé, O., Masson, S., et al. (2021). Inhibitory control and the understanding of buoyancy from childhood to adulthood. *J. Exp. Child Psychol.* 208:105155. doi: 10.1016/j.jecp.2021.105155

Brock, B., and Hunt, W. A. (1991). "Report on the formal specification and partial verification of the VIPER microprocessor," in *Proceedings of the Sixth Annual Conference on Computer Assurance: COMPASS '91*, (Gaithersburg, MD: IEEE), 91–98. doi: 10.1109/CMPASS.1991.161048

Buzsáki, G. (2010). Neural syntax: cell assemblies, synapsembles, and readers. *Neuron* 68, 362–385. doi: 10.1016/j.neuron.2010.09.023

Buzsáki, G., Anastassiou, C. A., and Koch, C. (2012). The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. *Nat. Rev. Neurosci.* 13, 407–420. doi: 10.1038/nrn3241

Cariani, P. A. (2009). The homeostat as embodiment of adaptive control. *Int. J. Gen. Syst.* 38, 139–154. doi: 10.1080/03081070802633593

Changeux, J. P., Courrège, P., and Danchin, A. (1973). A theory of the epigenesis of neuronal networks by selective stabilization of synapses. *Proc. Natl. Acad. Sci. U.S.A.* 70, 2974–2978. doi: 10.1073/pnas.70.10.2974

Changeux, J.-P. (2013). The concept of allosteric interaction and its consequences for the chemistry of the brain. *J. Biol. Chem.* 288, 26969–26986. doi: 10.1074/jbc.X113.503375

Changeux, J.-P., Goulas, A., and Hilgetag, C. C. (2021). A connectomic hypothesis for the hominization of the brain. *Cereb. Cortex* 31, 2425–2449. doi: 10.1093/cercor/bhaa365

Chaudhuri, R., and Fiete, I. (2016). Computational principles of memory. *Nat. Neurosci.* 19, 394–403. doi: 10.1038/nn.4237

Chen, I., and Lui, F. (2021). "Neuroanatomy, neuron action potential," in *StatPearls*, (Treasure Island, FL: StatPearls Publishing). Available online at: http://www.ncbi.nlm.nih.gov/books/NBK546639/ (accessed September 8, 2021).

Chung, A., Jou, C., Grau-Perales, A., Levy, E. R. J., Dvorak, D., Hussain, N., et al. (2021). Cognitive control persistently enhances hippocampal information processing. *Nature* 600, 484–488. doi: 10.1038/s41586-021-04070-5

Coavoux, M. (2021). "BERT-proof syntactic structures: investigating errors in discontinuous constituency parsing," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, (Stroudsburg, PA: Association for Computational Linguistics), 3259–3272. doi: 10.18653/v1/2021.findings-acl.288

Comiter, M. (2019). *Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It*. Available Online at: https://www.belfercenter.org/publication/AttackingAI (accessed March 23, 2022).

Conant, R. C., and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system †. *Int. J. Syst. Sci.* 1, 89–97. doi: 10.1080/00207727008920220

Copeland, B. J. (2020). "The Church-Turing thesis," in *The Stanford Encyclopedia of Philosophy*, (Metaphysics Research Lab, Stanford University). Available online at: https://plato.stanford.edu/archives/sum2020/entries/church-turing (accessed March 23, 2022).

Coriolis, G. (1836). Note sur un moyen de tracer des courbes données par des équations différentielles. *J. Math. Pures Appl.* I, 5–9.

Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd Edn. Hoboken, NJ: Wiley-Interscience.

Czech, J., Willig, M., Beyer, A., Kersting, K., and Fürnkranz, J. (2020). Learning to play the chess variant crazyhouse above world champion level with deep neural networks and human data. *Front. Artif. Intell.* 3:24. doi: 10.3389/frai.2020.00024

Danchin, A. (1998). The Delphic boat or what the genomic texts tell us. *Bioinformatics* 14:383. doi: 10.1093/bioinformatics/14.5.383

Danchin, A. (2002). Not every truth is good. The dangers of publishing knowledge about potential bioweapons. *EMBO Rep.* 3, 102–104. doi: 10.1093/embo-reports/kvf040

Danchin, A. (2009a). Bacteria as computers making computers. *FEMS Microbiol. Rev.* 33, 3–26. doi: 10.1111/j.1574-6976.2008.00137.x

Danchin, A. (2009b). Information of the chassis and information of the program in synthetic cells. *Syst. Synth. Biol.* 3, 125–134. doi: 10.1007/s11693-009-9036-5

Danchin, A. (2012). Scaling up synthetic biology: do not forget the chassis. *FEBS Lett.* 586, 2129–2137. doi: 10.1016/j.febslet.2011.12.024

Danchin, A. (2015). "The cellular chassis as the basis for new functionalities: shortcomings and requirements," in *Synthetic Biology* Risk Engineering, eds B. Giese, C. Pade, H. Wigger, and A. von Gleich (Cham: Springer International Publishing), 155–172. doi: 10.1007/978-3-319-02783-8_8

Danchin, A. (2021b). Three overlooked key functional classes for building up minimal synthetic cells. *Synth. Biol.* 6:ysab010. doi: 10.1093/synbio/ysab010

Danchin, A. (2021a). Biological innovation in the functional landscape of a model regulator, or the lactose operon repressor. *C R Biol.* 344, 111–126. doi: 10.5802/crbiol.52

Danchin, A., and Fang, G. (2016). Unknown unknowns: essential genes in quest for function. *Microb. Biotechnol.* 9, 530–540. doi: 10.1111/1751-7915.12384

Daniel, R., Rubens, J. R., Sarpeshkar, R., and Lu, T. K. (2013). Synthetic analog computation in living cells. *Nature* 497, 619–623. doi: 10.1038/nature12148

Darden, L., and Cain, J. A. (1989). Selection type theories. *Philos. Sci.* 56, 106–129. doi: 10.1086/289475

Dehaene, S., and Changeux, J. P. (2000). Reward-dependent learning in neuronal networks for planning and decision making. *Prog. Brain Res.* 126, 217–229. doi: 10.1016/S0079-6123(00)26016-0

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., and Gallant, J. L. (2019). The representation of semantic information across human cerebral cortex during listening versus reading is Invariant to stimulus modality. *J. Neurosci.* 39, 7722–7736. doi: 10.1523/JNEUROSCI.0675-19.2019

Dvorak, D., Chung, A., Park, E. H., and Fenton, A. A. (2021). Dentate spikes and external control of hippocampal function. *Cell Rep.* 36:109497. doi: 10.1016/j.celrep.2021.109497

Dvorak, D., Radwan, B., Sparks, F. T., Talbot, Z. N., and Fenton, A. A. (2018). Control of recollection by slow gamma dominating mid-frequency gamma in hippocampus CA1. *PLoS Biol.* 16:e2003354. doi: 10.1371/journal.pbio.2003354

Edelman, G. M. (1987). *Neural Darwinism: The Theory of Neuronal Group Selection*. New York, NY: Basic Books.

Edelman, G. M., and Gally, J. A. (2013). Reentry: a key mechanism for integration of brain function. *Front. Integr. Neurosci.* 7:63. doi: 10.3389/fnint.2013.00063

Edelman, G. M., Gally, J. A., and Baars, B. J. (2011). Biology of consciousness. *Front. Psychol.* 2:4. doi: 10.3389/fpsyg.2011.00004

Farini, D., Marazziti, D., Geloso, M. C., and Sette, C. (2021). Transcriptome programs involved in the development and structure of the cerebellum. *Cell. Mol. Life Sci*. 78, 6431–6451. doi: 10.1007/s00018-021-03911-w

Fenton, A. A. (2015). Excitation-inhibition discoordination in rodent models of mental disorders. *Biol. Psychiatry* 77, 1079–1088. doi: 10.1016/j.biopsych.2015.03.013

Fenton, A. A., Arolfo, M. P., Nerad, L., and Bures, J. (1995). Interhippocampal synthesis of lateralized place navigation engrams. *Hippocampus* 5, 16–24. doi: 10.1002/hipo.450050104

Freeth, T., Higgon, D., Dacanalis, A., MacDonald, L., Georgakopoulou, M., and Wojcik, A. (2021). A model of the cosmos in the ancient Greek Antikythera Mechanism. *Sci. Rep.* 11:5821. doi: 10.1038/s41598-021-84310-w

Gabbay, F., and Mendelson, A. (2021). Asymmetric aging effect on modern microprocessors. *Microelectron. Reliabil.* 119:114090. doi: 10.1016/j.microrel.2021.114090

Gardner, T. S., Cantor, C. R., and Collins, J. J. (2000). Construction of a genetic toggle switch in *Escherichia coli*. *Nature* 403, 339–342. doi: 10.1038/35002131

Girard, B., and Berthoz, A. (2005). From brainstem to cortex: computational models of saccade generation circuitry. *Prog. Neurobiol.* 77, 215–251. doi: 10.1016/j.pneurobio.2005.11.001

González-Acosta, C. A., Escobar, M. I., Casanova, M. F., Pimienta, H. J., and Buriticá, E. (2018). Von Economo neurons in the human medial frontopolar cortex. *Front. Neuroanat.* 12:64. doi: 10.3389/fnana.2018.00064

Hartree, D. R. (1940). The Bush differential analyser and its applications. *Nature* 146, 319–323. doi: 10.1038/146319a0

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science* 298, 1569–1579. doi: 10.1126/science.298.5598.1569

Hawkins, J., and Blakeslee, S. (2005). *On intelligence. How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines*, 1st Edn. New York, NY: Owl Books.

Hebb, D. O. (2002). *The Organization of Behavior: A Neuropsychological Theory*. Mahwah, NJ: L. Erlbaum Associates.

Hecker, A., Schulze, W., Oster, J., Richter, D. O., and Schuster, S. (2020). Removing a single neuron in a vertebrate brain forever abolishes an essential behavior. *Proc. Natl. Acad. Sci. U.S.A.* 117, 3254–3260. doi: 10.1073/pnas.1918578117

Hobert, O. (2021). Homeobox genes and the specification of neuronal identity. *Nat. Rev. Neurosci.* 22, 627–636. doi: 10.1038/s41583-021-00497-x

Hofstadter, D. R. (1999). *Gödel, Escher, Bach: An Eternal Golden Braid. 20. Anniversary*. New York, NY: Basic Books.

Hofstadter, D. R. (2007). *I am a Strange Loop*. New York, NY: Basic Books.

Hofstadter, D. R., Gödel, K., Escher, M. C., and Bach, J. S. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. New York, NY: Basic Books.

Hosoya, T. (2019). The basic repeating modules of the cerebral cortical circuit. *Proc. Jpn. Acad. Ser. B Phys. Biol. Sci.* 95, 303–311. doi: 10.2183/pjab.95.022

Hoyle, F. (1957). *Black Cloud*. The Hague: William Heinemann Ltd.

Hsieh, C., Tsokas, P., Grau-Perales, A., Lesburguères, E., Bukai, J., Khanna, K., et al. (2021). Persistent increases of PKMζ in memory-activated neurons trace LTP maintenance during spatial long-term memory storage. *Eur. J. Neurosci.* 54, 6795–6814. doi: 10.1111/ejn.15137

Hummels, K. R., and Kearns, D. B. (2020). Translation elongation factor P (EF-P). *FEMS Microbiol. Rev.* 44, 208–218. doi: 10.1093/femsre/fuaa003

Jaynes, J. (2000). *The Origin of Consciousness in the Breakdown of the Bicameral Mind. 1. Mariner Books*. Boston, MA: Houghton Mifflin.

Kari, L., and Rozenberg, G. (2008). The many facets of natural computing. *Commun. ACM* 51, 72–83. doi: 10.1145/1400181.1400200

Khilkevich, A., Zambrano, J., Richards, M.-M., and Mauk, M. D. (2018). Cerebellar implementation of movement sequences through feedback. *eLife* 7:e37443. doi: 10.7554/eLife.37443

Kim, J., and Augustine, G. J. (2021). Molecular layer interneurons: key elements of cerebellar network computation and behavior. *Neuroscience* 462, 22–35. doi: 10.1016/j.neuroscience.2020.10.008

Kimura, R., Kang, S., Takahashi, N., Usami, A., Matsuki, N., Fukai, T., et al. (2011). Hippocampal polysynaptic computation. *J. Neurosci.* 31, 13168–13179. doi: 10.1523/JNEUROSCI.1920-11.2011

Kohn, A. F. (1998). Effects of synaptic noise on a neuronal pool model with strong excitatory drive and recurrent inhibition. *Biosystems* 48, 113–121. doi: 10.1016/s0303-2647(98)00056-2

La Rosa, C., Cavallo, F., Pecora, A., Chincarini, M., Ala, U., Faulkes, C. G., et al. (2020). Phylogenetic variation in cortical layer II immature neuron reservoir of mammals. *eLife* 9:e55456. doi: 10.7554/eLife.55456

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM J. Res. Dev.* 3, 184–191.

Landauer, R. (1996). The physical nature of information. *Phys. Lett. A* 217, 188–193. doi: 10.1016/0375-9601(96)00453-7

Lehtiö, P., and Kohonen, T. (1978). Associative memory and pattern recognition. *Med. Biol.* 56, 110–116.

Levy, I., Hasson, U., and Malach, R. (2004). One picture is worth at least a million neurons. *Curr. Biol.* 14, 996–1001. doi: 10.1016/j.cub.2004.05.045

Lin, C., Sherathiya, V. N., Oh, M. M., and Disterhoft, J. F. (2020). Persistent firing in LEC III neurons is differentially modulated by learning and aging. *Elife* 9:e56816. doi: 10.7554/eLife.56816

Little, W. D., and Soudack, A. C. (1965). On the analog computer solution of first-order partial differential equations. *Math. Comput. Simul.* 7, 190–194. doi: 10.1016/S0378-4754(65)80035-0

Lundgren, B. (2019). Does semantic information need to be truthful? *Synthese* 196, 2885–2906. doi: 10.1007/s11229-017-1587-5

Martel, J. N. P., Muller, L. K., Carey, S. J., Dudek, P., and Wetzstein, G. (2020). Neural sensors: learning pixel exposures for HDR imaging and video compressive sensing with programmable sensors. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 1642–1653. doi: 10.1109/TPAMI.2020.2986944

Mashour, G. A., Roelfsema, P., Changeux, J.-P., and Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105, 776–798. doi: 10.1016/j.neuron.2020.01.026

Mehta, M. R. (2015). From synaptic plasticity to spatial maps and sequence learning: place Field Plasticity. *Hippocampus* 25, 756–762. doi: 10.1002/hipo.22472

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. *Eur. J. Neurosci.* 48, 2609–2621. doi: 10.1111/ejn.13748

Michiels van Kessenich, L., Luković, M., de Arcangelis, L., and Herrmann, H. J. (2018). Critical neural networks with short- and long-term plasticity. *Phys. Rev. E* 97:032312. doi: 10.1103/PhysRevE.97.032312

Miłkowski, M. (2021). Correspondence: theory of semantic information. *Br. J. Philos. Sci.* 714804. doi: 10.1086/714804

Miller, E. K., and Buschman, T. J. (2007). "Rules through recursion: how interactions between the frontal cortex and basal ganglia may build abstract, complex rules from concrete, simple ones," in *Neuroscience of Rule-Guided Behavior*, eds S. A. Bunge and J. D. Wallis (Oxford: Oxford University Press), 419–440. doi: 10.1093/acprof:oso/9780195314274.003.0022

Miller, E. K., and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* 24, 167–202. doi: 10.1146/annurev.neuro.24.1.167

Misa, T. J. (2007). Understanding "How computing has changed the world.". *IEEE Ann. Hist. Comput.* 29, 52–63. doi: 10.1109/MAHC.2007.4407445

Mitchell, A. S. (2015). The mediodorsal thalamus as a higher order thalamic relay nucleus important for learning and decision-making. *Neurosci. Biobehav. Rev.* 54, 76–88. doi: 10.1016/j.neubiorev.2015.03.001

Mittal, D., and Narayanan, R. (2021). Resonating neurons stabilize heterogeneous grid-cell networks. *Elife* 10:e66804. doi: 10.7554/eLife.66804

Modgil, S., and Modgil, C. (eds). (1987). *Noam Chomsky: Consensus and Controversy*. New York, NY: Falmer Press.

Muzio, G., O'Bray, L., and Borgwardt, K. (2021). Biological network analysis with deep learning. *Brief. Bioinform.* 22, 1515–1530. doi: 10.1093/bib/bbaa257

Nadel, L., and Buresova, O. (1968). Monocular input and interhemispheric transfer in the reversible split-brain. *Nature* 220, 914–915. doi: 10.1038/220914a0

O'Leary, D. D. M., Chou, S.-J., and Sahara, S. (2007). Area patterning of the mammalian cortex. *Neuron* 56, 252–269. doi: 10.1016/j.neuron.2007.10.010

Pinto, M. A., Rosso, O. A., and Matias, F. S. (2019). Inhibitory autapse mediates anticipated synchronization between coupled neurons. *Phys. Rev. E* 99:062411. doi: 10.1103/PhysRevE.99.062411

Prifti, E., Chevaleyre, Y., Hanczar, B., Belda, E., Danchin, A., Clément, K., et al. (2020). Interpretable and accurate prediction models for metagenomics data. *Gigascience* 9:giaa010. doi: 10.1093/gigascience/giaa010

Quiroga, R. Q., Kreiman, G., Koch, C., and Fried, I. (2008). Sparse but not 'Grandmother-cell' coding in the medial temporal lobe. *Trends Cogn. Sci.* 12, 87–91. doi: 10.1016/j.tics.2007.12.003

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature* 435, 1102–1107. doi: 10.1038/nature03687

Roell, M., Viarouge, A., Hilscher, E., Houdé, O., and Borst, G. (2019). Evidence for a visuospatial bias in decimal number comparison in adolescents and in adults. *Sci. Rep.* 9:14770. doi: 10.1038/s41598-019-51392-6

Rosen, K. H. (2011). *Elementary Number Theory and its Applications*, 6. Edn. Boston, MA: Pearson.

Ryle, G. (2009). *The Concept of Mind*. London: Routledge.

Sahara, S., Yanagawa, Y., O'Leary, D. D. M., and Stevens, C. F. (2012). The fraction of cortical GABAergic neurons is constant from near the start of cortical neurogenesis to adulthood. *J. Neurosci.* 32, 4755–4761. doi: 10.1523/JNEUROSCI.6412-11.2012

Sakmann, B. (2017). From single cells and single columns to cortical networks: dendritic excitability, coincidence detection and synaptic transmission in brain slices and brains. *Exp. Physiol.* 102, 489–521. doi: 10.1113/EP085776

Schlör, D., Ring, M., and Hotho, A. (2020). iNALU: improved neural arithmetic logic unit. *Front. Artif. Intell.* 3:71. doi: 10.3389/frai.2020.00071

Schomburg, E. W., Fernández-Ruiz, A., Mizuseki, K., Berényi, A., Anastassiou, C. A., Koch, C., et al. (2014). Theta phase segregation of input-specific gamma patterns in entorhinal-hippocampal networks. *Neuron* 84, 470–485. doi: 10.1016/j.neuron.2014.08.051

Schuman, B., Dellal, S., Prönneke, A., Machold, R., and Rudy, B. (2021). Neocortical layer 1: an elegant solution to top-down and bottom-up integration. *Annu. Rev. Neurosci.* 44, 221–252. doi: 10.1146/annurev-neuro-100520-012117

Seger, C. A. (2006). The basal ganglia in human learning. *Neuroscientist* 12, 285–290. doi: 10.1177/1073858405285632

Shamir, I., and Assaf, Y. (2021). An MRI-based, data-driven model of cortical laminar connectivity. *Neuroinformatics* 19, 205–218. doi: 10.1007/s12021-020-09491-7

Sikorski, R. (1969). *Boolean algebras*. New York, NY: Springer-Verlag.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 1140–1144. doi: 10.1126/science.aar6404

Simon, H. A. (1974). *The Sciences of the Artificial. 4. Print*. Cambridge, MA: MIT Press.

Simon, H. A. (1991). "The architecture of complexity," in *Facets of Systems Science*, ed. G. E. Mobus (Boston, MA: Springer US), 457–476. doi: 10.1007/978-1-4899-0718-9_31

Soma, Y., Takahashi, M., Fujiwara, Y., Shinohara, T., Izumi, Y., Hanai, T., et al. (2021). Design of synthetic quorum sensing achieving induction timing-independent signal stabilization for dynamic metabolic engineering of *E. coli*. *ACS Synth. Biol.* 10, 1384–1393. doi: 10.1021/acssynbio.1c00008

Tecuatl, C., Wheeler, D. W., Sutton, N., and Ascoli, G. A. (2021). Comprehensive estimates of potential synaptic connections in local circuits of the rodent hippocampal formation by axonal-dendritic overlap. *J. Neurosci.* 41, 1665–1683. doi: 10.1523/JNEUROSCI.1193-20.2020

Thom, R., Thom, R., and Thom, R. (1990). *Semio Physics: A Sketch*. Redwood City, CA: Addison-Wesley.

Tsuda, B., Tye, K. M., Siegelmann, H. T., and Sejnowski, T. J. (2020). A modeling framework for adaptive lifelong learning with transfer and savings through gating in the prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 117, 29872–29882. doi: 10.1073/pnas.2009591117

Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungs problem. *Proc. Lond. Math. Soc.* s2–42, 230–265. doi: 10.1112/plms/s2-42.1.230

Vedral, V. (2012). *Decoding Reality: The Universe as Quantum Information. 1. Publ. in Paperback*. Oxford: Oxford University Press.

Wagstyl, K., Larocque, S., Cucurull, G., Lepage, C., Cohen, J. P., Bludau, S., et al. (2020). BigBrain 3D atlas of cortical layers: cortical and laminar thickness gradients diverge in sensory and motor cortices. *PLoS Biol.* 18:e3000678. doi: 10.1371/journal.pbio.3000678

Watanabe, S., Shiwa, Y., Itaya, M., and Yoshikawa, H. (2012). Complete sequence of the first chimera genome constructed by cloning the whole genome of *Synechocystis* strain PCC6803 into the *Bacillus subtilis* 168 genome. *J. Bacteriol.* 194:7007. doi: 10.1128/JB.01798-12

Wijekoon, J. H. B., and Dudek, P. (2012). VLSI circuits implementing computational models of neocortical circuits. *J. Neurosci. Methods* 210, 93–109. doi: 10.1016/j.jneumeth.2012.01.019

Wilts, B. D., Michielsen, K., De Raedt, H., and Stavenga, D. G. (2014). Sparkling feather reflections of a bird-of-paradise explained by finite-difference time-domain modeling. *Proc. Natl. Acad. Sci. U.S.A.* 111, 4363–4368. doi: 10.1073/pnas.1323611111

Zetterberg, L. H., Kristiansson, L., and Mossberg, K. (1978). Performance of a model for a local neuron population. *Biol. Cybern.* 31, 15–26. doi: 10.1007/BF00337367

Zhang, M., Lu, M., Huang, H., Liu, X., Su, H., and Li, H. (2019). Maturation of thalamocortical synapses in the somatosensory cortex depends on neocortical AKAP5 expression. *Neurosci. Lett.* 709:134374. doi: 10.1016/j.neulet.2019.134374

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

OPINION
published: 22 July 2021
doi: 10.3389/fcomp.2021.681416

The Brain-As-Computer Metaphor

Martin Davis[1,2]*

[1]*Independent Researcher, Berkeley, CA, United States,* [2]*New York University, New York City, NY, United States*

Keywords: brain, consciousness, algorithm, neural net, operating system

> He thought he saw an elephant
> That practiced on a fife
> He looked again, and found it was
> A letter from his wife
>
> –Lewis Carroll[1]

> I write as someone who is old (I'm approaching my 93rd birthday) and has a brain. While I claim no expertise in brain science, I hope to suggest questions that might occur to a computer scientist thinking about the brain. What little I have learned about the brain comes from a few books (Patricia, 1986; Dennett, 1991; Hobson, 1994). I have also benefited from lectures by Patricia Churchland, one of the authors.

Having an old brain is not wonderful; I can't help but be aware of how much of what it could do when I was twenty is gone. It is so slow. However, this does provide one advantage: Various stages of processes that my brain carries out, that in past years would have gone by too quickly for me to have noticed them, are now quite evident.[2]

1 DOES THE BRAIN USE ALGORITHMS?

I ask my friend: "Have you ever read anything by Turgenev?" Her negative reply comes with no pause. Does her brain have a database of fiction she has read? Is her brain using a search algorithm? If not, how else can we imagine this feat accomplished?

I ask: "What's the name of the man you were seated next to at dinner last night?" She doesn't remember. A half hour later, while we were talking about something different, she says"I remember now, Jerome's his name. It just popped into my head." If the brain used a search algorithm to do this, might it be different from that of the previous example, a slower, but more methodical, procedure?

Can we imagine a device made from the brain's "hardware" that can execute search algorithms? Or arbitrary algorithms for that matter? In my (Davis, 2017), I emphasized that very little is needed for Turing completeness. No doubt a universal computer could be built with the brain's neurons. However, it's much less clear that one could evolve. The case of the genetic code in which amino acids are coded by strings suggests that the possibility is not so far fetched. Also in fact, spoken and written language are examples of arbitrary symbols representing objects, actions, and concepts.

Someone crosses a busy street, expertly weaving among the cars. How would one program a robot to do this? Until recently, a process using much numerical computation would be proposed. Nowadays, one could consider the alternative of "training" a multi-level neural net for the purpose. It is certainly easier to imagine brains doing something like this than carrying out a process involving a lot of arithmetic computation.

OPEN ACCESS

Edited by:
Bud (Bhubaneswar) Mishra,
New York University, United States

Reviewed by:
Saket Navlakha,
Cold Spring Harbor Laboratory,
United States
Jantine Broek,
Imperial College London,
United Kingdom

****Correspondence:***
Martin Davis
martin@eipye.com

Specialty section:
This article was submitted to
Theoretical Computer Science,
a section of the journal
Frontiers in Computer Science

Received: *25 March 2021*
Accepted: *26 April 2021*
Published: *22 July 2021*

Citation:
Davis M (2021) The Brain-As-Computer Metaphor.
Front. Comput. Sci. 3:681416.
doi: 10.3389/fcomp.2021.681416

[1] Carroll, 1991

[2] I am grateful for the comments of the two anonymous reviewers.

A student is studying calculus, specifically integration. A bunch of techniques are learned: change of variable, integration by parts, partial fractions. How to decide which to use when presented with a specific example? On the basis of their own attempts as well as access to worked-out examples, students can develop an intuition that guides them to the best choice of technique. Is the student's training like that of a neural net?

2 VISION AND OTHER BRAIN PRODUCTS

Scientists studying human vision have shown that what we "see" is the result of complex data processing by the brain. Continuous rapid eyeball motions send huge amounts of data to the brain from which the brain produces the scene presented to us. Our sense that we just see that which is before us is a brain product. We actually "see" a sequence of converging edits (as Lewis Carroll playfully suggests in the quoted excerpt in the heading). With my slow old brain doing the work, early edits are sometimes of sufficient duration that I "see" quite clearly something that isn't there, before the corrected version appears.

How much of our inner mental experience is similarly illusory? Eighty years ago, I was approaching my thirteenth birthday. I remember many things about my life and experiences at that earlier time, but as we have learned, these memories may be unreliable. I have a strong sense that the old man writing this and that boy are different stages in life of the very same person. But isn't that very sense also a brain product?

3 DOES THE BRAIM HAVE AN OPERATING SYSTEM?

Many years ago, there was a candy vending machine where I worked. For health reasons, I needed to avoid being tempted by its offerings. One day, passing the machine on my way from my office to the men's room. I was strongly tempted because one of my favorites, chocolate with almonds, had just become available. I resolved to resist the temptation and thought no more about it. As I walked back to my office from the toilet, thinking about a mathematical problem, I noticed that I was eating that delicious candy bar with no memory of having bought it.

This story can be conceptualized as a struggle among three brain processes, we may call: EatTheSweet, EatHealthy, and DoMath. EatHealthy executes and stops EatTheSweet. Then later, while DoMath is going at full blast, EatTheSweet sees its chance and executes. What controls this? In a computer, it would be the operating system that allocates resources to processes and permits them to execute. One can certainly imagine that the multitasking brain possesses some such mechanism controlling its hundreds of processes, some struggling for attention and resources. But an operating system needs a user interface. And where and what is the user?

Our sensation of consciousness and, in particular, our I-me sense of ourself as an individual, have presumably evolved because having them provides an evolutionary advantage. I suggest that functionally, consciousness serves as an interface to the brain's operating system. And furthermore, that the I-me sense, perhaps the most remarkable brain product, functions as the user.

AUTHOR CONTRIBUTIONS

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

REFERENCES

Carroll, L. (1991). *The Complete Sylvie and Bruno*. San Francisco: Mercury House.

Patricia, S. C. (1986). *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge: MIT Press.

Davis, M. (2017). "Universality Is Ubiquitous," in *Philosophical Explorations of the Legacy of Alan Turing*. Editors J Floyd and A Bokulich (Springer), 153–158. doi:10.1007/978-3-319-53280-6_6

Dennett, D. (1991). *Consciousness Explained*. Brown: Little.

Hobson, J. A. (1994). *The Chemistry of Conscious States*. Brown: Little.

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

The handling editor declared a shared affiliation with with the author at time of review.

frontiers
in Ecology and Evolution

PERSPECTIVE
published: 18 October 2021
doi: 10.3389/fevo.2021.755981

Intelligence as Information Processing: Brains, Swarms, and Computers

Carlos Gershenson [1,2,3]*

[1] Departamento de Ciencias de la Computación, Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas, Universidad Nacional Autónoma de México, Mexico City, Mexico, [2] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico, [3] Lakeside Labs GmbH, Klagenfurt am Wörthersee, Austria

There is no agreed definition of intelligence, so it is problematic to simply ask whether brains, swarms, computers, or other systems are intelligent or not. To compare the potential intelligence exhibited by different cognitive systems, I use the common approach used by artificial intelligence and artificial life: Instead of studying the substrate of systems, let us focus on their organization. This organization can be measured with information. Thus, I apply an informationist epistemology to describe cognitive systems, including brains and computers. This allows me to frame the usefulness and limitations of the brain-computer analogy in different contexts. I also use this perspective to discuss the evolution and ecology of intelligence.

Keywords: mind, cognition, intelligence, information, brain, computer, swarm

OPEN ACCESS

Edited by:
Giorgio Matassi,
FRE3498 Ecologie et Dynamique des Systèmes Anthropisés (EDYSAN), France

Reviewed by:
Thilo Gross,
Helmholtz Institute for Functional Marine Biodiversity (HIFMB), Germany
Alberto Policriti,
University of Udine, Italy

****Correspondence:***
Carlos Gershenson
cgg@unam.mx

Specialty section:
This article was submitted to Models in Ecology and Evolution, a section of the journal Frontiers in Ecology and Evolution

Received: *09 August 2021*
Accepted: *22 September 2021*
Published: *18 October 2021*

Citation:
Gershenson C (2021) Intelligence as Information Processing: Brains, Swarms, and Computers. Front. Ecol. Evol. 9:755981. doi: 10.3389/fevo.2021.755981

1. INTRODUCTION

In the 1850s, an English newspaper described the growing global telegraph network as a "nervous system of the planet" (Gleick, 2011). Notice that this was half a century before Ramón y Cajal (1899) first published his studies on neurons. Still, metaphors have been used since antiquity to describe and try to understand our bodies and our minds (Zarkadakis, 2015; Epstein, 2016): humans have been described as made of clay (Middle East) or corn (Americas), with flowing humors, like clockwork automata, similar to industrial factories, etc. The most common metaphor in cognitive sciences has been that of describing brains as computers (von Neumann, 1958; Davis, 2021).

Metaphors have been used in a broad range of disciplines. For example, in urbanism, there are arguments in favor of changing the dominant narrative of "cities as machines" to "cities as organisms" (Batty, 2012; Gershenson, 2013b).

We can have a plethora of discussions on which metaphors are the best. Still, being pragmatic, we can judge metaphors in terms of their usefulness: if they help us understand phenomena or build systems, then they are valuable. Notice that then, depending on the context, different metaphors can be useful for different purposes (Gershenson, 2004). For example, in the 1980s, the debate between symbolists/representationists (brain as processing symbols) (Fodor and Pylyshyn, 1988) and connectionists (brain as network of simple units) (Smolensky, 1988) did not end with a "winner" and a "loser," as both metaphors (computational, by the way) are useful in different contexts.

There have been several other metaphors used to describe cognition, minds, and brains, each with their advantages and disadvantages (Varela et al., 1991; Steels and Brooks, 1995; Clark and Chalmers, 1998; Beer, 2000; Gärdenfors, 2000; Garnier et al., 2007; Chemero, 2009; Froese and Ziemke, 2009; Kiverstein and Clark, 2009; Froese and Stewart, 2010; Stewart et al., 2010; Downing, 2015; Harvey, 2019). It is not my purpose to discuss these here, but to notice that there is a rich variety of flavors when it comes to studying cognition. Nevertheless, all of these metaphors can be described in terms of information processing. Since computation can be understood as the transformation of information (Gershenson, 2012), "computers," broadly understood as machines that process information can be a useful metaphor to contain and compare other metaphors. Note that the concept of "machine" (and thus computer) could also be updated (Bongard and Levin, 2021).

Formally, computation was defined by Turing (1937). A computable function is that which can be calculated by a Universal Turing Machine (UTM). Still, there are two main limitations of UTMs related to modeling minds (Gershenson, 2011a):

1. **UTMs are closed**. Once a computation begins, there is no change in the program or data, so adaptation during computation is limited.
2. **UTMs compute only once they halt**. In other words, outputs depend on a UTM "finishing its computation." Still, minds seem to be more continuous than halting. Then the question arises: what function would a mind be computing?

As many have noted, the continuous nature of cognition seems to be closely related to that of the living (Maturana and Varela, 1980; Hopfield, 1994; Stewart, 1995; Walker, 2014). We have previously studied the "living as information processing" (Farnsworth et al., 2013), not only at the organism level, but at all relevant scales. Thus, it is natural to use a similar approach to describe intelligence.

Note that the limitations of UTMs apply only for *theoretical* computation. In practice, many artificial computation systems are continuous, such as reactive systems. An example would be an operating system, that does not precisely halt, but is always expecting events (internal or external) and responding to these.

In the next section, I present a general notion of information and its limits to study intelligence. Then, I present the advantages of studying intelligence in terms of information processing. Intelligence is not restricted to brains, and swarms are a classic example of this, which can also be described as information processing systems. Before concluding, I exploit the metaphor of "intelligence as information processing" to understand its evolution and ecology.

2. INFORMATION

Shannon (1948) proposed a measure of information in the context of telecommunications, that is equivalent to Boltzmann-Gibbs entropy. This measure characterizes how much a receiver "learns" from incoming symbols (usually bits) of a string, based on the probability distribution of previously known/received symbols: if new bits can be completely determined from the past (as in a string with only one repeating symbol), then they carry zero information (because we know that the new symbols will be the same as previous ones). If previous information is useless to predict the next bit (as in a random coin toss), then the bit will carry maximum information. Elaborating on this, Shannon calculated how much redundancy is required to reliably transmit a message over an unreliable (noisy) channel. Even when Shannon's purpose was very specific, the use of information in various disciplines has exploded in recent decades (Haken, 1988; Lehn, 1990; Wheeler, 1990; Gell-Mann and Lloyd, 1996; Atlan and Cohen, 1998; DeCanio and Watkins, 1998; Roederer, 2005; von Baeyer, 2005; Cover and Thomas, 2006; Prokopenko et al., 2009, 2011; Batty et al., 2012; Escalona-Morán et al., 2012; Gershenson, 2012, 2020, 2021b; Fernández et al., 2014, 2017; Zubillaga et al., 2014; Haken and Portugali, 2015; Hidalgo, 2015; Murcio et al., 2015; Amoretti and Gershenson, 2016; Roli et al., 2018; Equihua et al., 2020; Krakauer et al., 2020; Scharf, 2021).

We can say that electronic computers process information *explicitly*, as we can analyze each change of state and information is encoded in a precise physical location. However, humans and other animals process information *implicitly*. For example, we say we have memories, but these are not physically at a specific location. And it seems unfeasible to represent precisely the how information changes in our brains. Still, we do process information, as we can describe "inputs" (perceptions) and "outputs" (actions).

Shannon assumed that the *meaning* of a message was agreed previously between emitter and receiver. This was no major problem for telecommunications. However, in other contexts, meaning is not a trivial matter. Following Wittgenstein (1999), we can say that the meaning of information is given by the *use* agents make of it. This has several implications. One is that we can change meaning without changing information [passive information transformation; (Gershenson, 2012)]. Another is the limits on artificial intelligence (Searle, 1980; Mitchell, 2019), as the *use* of information in artificial systems tends to be predefined. Algorithms can "recognize" traffic lights or cats in an image, as they are trained for this specific purpose. But the "meaning" for computer programs is predefined, i.e., what we want the program to do. The quest for an "artificial general intelligence" that would go beyond this limit has produced not much more than speculations.

Even if we could simulate in a digital computer all the neurons, molecules, or even elementary particles from a brain, such a simulation would not yield something akin to a mind. On the one hand, *interactions* generate novel information at multiple scales, so we would need to include not only brain, but body and world that interacts with the brain (Clark, 1997). Moreover, such a simulation would require to model not only one scale, but all scales relevant to minds (see below). On the other hand, as mentioned above, *observers* can give different meanings to the same information. In other words, the same "brain state" for different people could refer to different "mental states." For example, we could use the same simple "neural" architecture of a Braitenberg vehicle (Braitenberg, 1986) that exhibits phototaxis, but connect the inputs to different sensors (e.g., sound or odor, instead of light), and the "meaning" of the information processed by the same neural architecture would be very different. In a sense, this is related to the failure of Laplace's daemon: even with full information of the states of the components of a system, prediction is limited because interactions generate novel information (Gershenson, 2013a). And this novel information can determine the future production of information at different scales through upward or downward causation (Campbell, 1974; Bitbol, 2012; Farnsworth et al., 2017; Flack, 2017), so all relevant scales should be considered (Gershenson, 2021a). An example of downward causation can be given with money: it is a social contract, but has a causal effect on matter and energy (physics), e.g., when we extract minerals from a mountain. This action does not violate the laws of physics, but the laws of physics are not enough to predict that the matter in the mountain will be extracted by humans for their own purposes.

In spite of all its limitations, the computer metaphor can be useful in a particular way. First, the limits on prediction by interactions are related to *computational irreducibility* (Wolfram, 2002). Second, describing brains and minds in terms of information allows us to avoid dualisms. Thus, it becomes natural to use information processing to describe intelligence and its evolution. Finally, information can contain other metaphors and formalisms, so it can be used to compare them and also to exploit their benefits.

3. INTELLIGENCE

There are several definitions of intelligence, but not a single one that is agreed upon. We have similar situations with the definitions of life (De Duve, 2003; Aguilar et al., 2014), consciousness (Michel et al., 2019), complexity (Lloyd, 2001; Heylighen et al., 2007), emergence (Bedau and Humphreys, 2008), and more. These concepts could be said to be of the type "I know it when I see it," to quote Potter Stewart.

Still, having no agreed definition is no motive nor excuse for not studying a phenomenon. Moreover, having different definitions for the same phenomenon can give us broader insights than if we stick to a single, narrow, inflexible definition.

Thus, we could define intelligence as "the art of getting away with it" (Arturo Frappé), or "the ability to hold two opposed ideas in mind at the same time and still retain the ability to function. One should, for example, be able to see that things are hopeless and yet be determined to make them otherwise" (F. Scott Fitzgerald). Turing (1950) proposed his famous test to decide whether a machine was intelligent. Generalizing Turing's test, Mario Lagunez suggested that in order to decide whether a system was intelligent, first, the system has to perform an action. Then, an observer has to *judge* whether the action was intelligent or not, according to some criteria. In this sense, there is no intrinsically intelligent behavior. All actions and decisions are contextual (Gershenson, 2002). Like with meaning, the same action can be intelligent or not, depending on the context and on the judge and their expectations.

Generalizing, we can define intelligence in terms of information processing: An agent a can be described as intelligent if it transforms information [individual (internal) or environmental (external)] to increase its "satisfaction" σ.

I have previously defined satisfaction $\sigma \in [0, 1]$ as the degree to which the goals of an agent have been fulfilled (Gershenson, 2007, 2011b). Certainly, we still require an observer, since we are the ones who define the goals of an agent, its boundaries, its scale, and thus, its satisfaction. Examples of goals are sustainability, survival, happiness, power, control, and understanding. All of these can be described as *information propagation* (Gershenson, 2012): In this context, an intelligent agent will propagate its own information.

Brains by themselves cannot propagate. But species of animals with brains tend to propagate. In this context, brains are parts of agents that help process information in order to propagate those agents. From this abstract perspective, we can see that such ability is not restricted to brains (Levin and Dennett, 2020). Thus, there are other mechanisms capable of producing intelligent behavior.

4. SWARMS

There has been much work related to collective intelligence and cognition (Hutchins, 1995; Heylighen, 1999; Reznikova, 2007; Couzin, 2009; Malone and Bernstein, 2015; Solé et al., 2016). Interestingly, groups of humans, animals or machines do not have a single brain. Thus, information processing is distributed.

A particular case is that of insect swarms (Chialvo and Millonas, 1995; Garnier et al., 2007; Passino et al., 2008; Marshall et al., 2009; Trianni and Tuci, 2009; Martin and Reggia, 2010), where not only information processing is distributed, but also reproduction and selection occur at the colony level (Hölldobler and Wilson, 2008).

To compare the cognitive architectures of brains and swarms, I previously proposed *computing networks* (Gershenson, 2010). With this formalism, it can be shown that the differences in substrate do not necessarily imply a theoretical difference in cognitive abilities. Nevertheless, in practice, the speed and scalability of information processing of brains is much superior than that of swarms: neurons can interact in the scale of milliseconds, and mammal brains can have a number of neurons in the order of 10^{11} with 10^{14} synapses (several species have more neurons than humans, including elephants and some whales, orcas having the most and more than twice as humans). The largest insect swarms that have been registered (locusts) are also in the order of 10^{11} individuals (covering $200 Km^2$). However, insects interact in the scale of seconds, and only with their local neighbors. In theory, it might not matter much. But in practice, this limits considerably the information processing capacities of swarms over brains.

Thus, the brain as computer metaphor is not appropriate for studying collective intelligence in general, nor swarm intelligence in particular. However, the intelligence of brains and swarms can be described in terms of information processing, as an agent a can be an organism or a colony, with its own satisfaction σ defined by an external observer.

Another advantage of studying intelligence as information processing is that we can use the same formalism to study intelligence at multiple scales: cellular, multicellular, collective/social, and cultural. Curiously, at the global scale (where we might reach a scale of 10^{11} humans later this century), the brain metaphor has also been used (Mayer-Kress and Barczys, 1995; Börner et al., 2005; Bernstein et al., 2012), although its usefulness remains to be demonstrated.

5. EVOLUTION AND ECOLOGY

If we want to have a better understanding of intelligence, we must study how it came to evolve. Intelligence as information-processing can also be useful in this context, as different substrates and mechanisms can be used to exhibit intelligent behavior.

What could be the ecological pressures that promote the evolution of intelligence? Since environments and ecosystems can also be described in terms of information, we can say that more *complex* environments will promote—through natural selection—more complex organisms and species, which will require a more complex intelligence to process the information of their environment and of other organisms and species they interact with (Gershenson, 2012). In this way, the complexity of ecosystems can also be expected to increase though evolution. It should be noted that we understand complexity as a balance between order and chaos, stability and change (Packard, 1988; Langton, 1990; Lopez-Ruiz et al., 1995; Fernández et al., 2014; Roli et al., 2018). Thus, species cannot be too robust or too adaptable in order to thrive in a complex ecosystem. This certainly will depend on how stable or volatile the ecosystems will be Equihua et al. (2020), but it is clear that organisms require to match the *variety* that their environment poses (Ashby, 1956; Gershenson, 2015) (see below).

These ideas generalize Dunbar's (1993, 2003) "social brain hypothesis": larger and more complex social groups put a selective pressure on more complex information processing (measured as the neocortex to bodymass ratio), which gives individuals more cognitive capacities to recognize different individuals, remember who can they trust, multiple levels of intentionality (Dennett, 1989), and so on. In turn, increased cognitive abilities lead to more complex groups, so this cycle reinforces the selection for more intelligent individuals.

One can make a similar argument using environments instead of social groups: more complex ecosystems put a selective pressure for more intelligent organisms, social groups, and species; as they require greater information-processing capabilities to survive and exploit their environments. This also creates a feedback, where more complex information processing by organisms, groups, and species produce more complex ecosystems.

However, individuals can "offload" their information processing to their group or environment, leading to a decrease in their individual information processing abilities (Reséndiz-Benhumea et al., 2021). This is to say that intelligence does not always increase. Although there is a selective pressure for intelligence, its cost imposes limits that depend as well on the usefulness of increased cognitive abilities.

Generalizing, we can say that information evolves to have greater control over its own production (Gershenson, 2012). This leads to more complex information-processing, and thus, we can expect intelligence to increase at multiple scales through evolution, independently on the substrates that actually do the information processing.

Another way of describing the same: information is transformed by different causes. This generates a *variety* of complexity (Ashby, 1956; Gershenson, 2015). More complex information requires more complex agents to propagate it, leading to an increase of complexity and intelligence through evolution.

At different scales, since the Big Bang, we have seen an increase of information processing through evolution. In recent decades, this increase has been supraexponential in computers (Schaller, 1997). Although there are limitations for sustaining this rate of increase (Shalf, 2020), we can say that the increase of intelligence is a natural tendency of evolution, be it of brains, swarms, or machines. This will not lead to a "singularity," but to an increase of the intelligence and complexity of humans, machines, and the ecosystems we create.

6. CONCLUSION

Brains are not essential for intelligence. Plants, swarms, bacterial colonies, robots, societies, and more exhibit intelligence without brains. An understanding of intelligence (and life, Gershenson et al., 2020) independently of its substrate, in terms of information processing, will be more illuminating that focussing only on the mechanisms used by vertebrates and other animals. In this sense, the metaphor of the brain as a computer, is limited more on the side of the brain than on the side of the computer. Brains do process information to exhibit intelligence, but there are several other mechanisms that also process information to exhibit intelligence. Brains are just a particular case, and we can learn a lot from them, but we will learn more if we do not limit our studies to their particular type of cognition.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

CG conceived and wrote the paper.

FUNDING

This work was supported by UNAM's PAPIIT IN107919 and IV100120 grants.

REFERENCES

Aguilar, W., Santamaría Bonfil, G., Froese, T., and Gershenson, C. (2014). The past, present, and future of artificial life. *Front. Robot. AI* 1:8. doi: 10.3389/frobt.2014.00008

Amoretti, M., and Gershenson, C. (2016). Measuring the complexity of adaptive peer-to-peer systems. *Peer-to-Peer Netw. Appl*. 9, 1031–1046. doi: 10.1007/s12083-015-0385-4

Ashby, W. R. (1956). *An Introduction to Cybernetics*. London: Chapman & Hall. doi: 10.5962/bhl.title.5851

Atlan, H., and Cohen, I. R. (1998). Immune information, self-organization and meaning. *Int. Immunol*. 10, 711–717. doi: 10.1093/intimm/10.6.711

Batty, M. (2012). Building a science of cities. *Cities* 29, S9–S16. doi: 10.1016/j.cities.2011.11.008

Batty, M., Morphet, R., Massuci, P., and Stanilov, K. (2012). "Entropy, complexity and spatial information," in *CASA Working Paper, 185*. London, UK.

Bedau, M. A., and Humphreys, P. (eds.). (2008). *Emergence: Contemporary Readings in Philosophy and Science*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262026215.001.0001

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends Cogn. Sci*. 4, 91–99. doi: 10.1016/S1364-6613(99)01440-0

Bernstein, A., Klein, M., and Malone, T. W. (2012). Programming the global brain. *Commun. ACM* 55, 41–43. doi: 10.1145/2160718.2160731

Bitbol, M. (2012). Downward causation without foundations. *Synthese* 185, 233–255. doi: 10.1007/s11229-010-9723-5

Bongard, J., and Levin, M. (2021). Living things are not (20th century) machines: updating mechanism metaphors in light of the modern science of machine behavior. *Front. Ecol. Evol*. 9:147. doi: 10.3389/fevo.2021.650726

Börner, K., Dall'Asta, L., Ke, W., and Vespignani, A. (2005). Studying the emerging global brain: analyzing and visualizing the impact of co-authorship teams. *Complexity* 10, 57–67. doi: 10.1002/cplx.20078

Braitenberg, V. (1986). *Vehicles: Experiments in Synthetic Psychology*. Cambridge, MA: MIT Press.

Campbell, D. T. (1974). "'Downward causation' in hierarchically organized biological systems," in *Studies in the Philosophy of Biology*, eds F. J. Ayala and T. Dobzhansky (New York City, NY: Macmillan), 179–186. doi: 10.1007/978-1-349-01892-5_11

Chemero, A. (2009). *Radical Embodied Cognitive Science*. Cambridge, MA: The MIT Press. doi: 10.7551/mitpress/8367.001.0001

Chialvo, D., and Millonas, M. (1995). "How swarms build cognitive maps," in *The Biology and Technology of Intelligent Autonomous Agents, Vol. 144*, ed L. Steels (Berlin; Heidelberg: Springer), 439–450. doi: 10.1007/978-3-642-79629-6_20

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/1552.001.0001

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis* 58, 7–19. doi: 10.1093/analys/58.1.7

Couzin, I. D. (2009). Collective cognition in animal groups. *Trends Cogn. Sci*. 13, 36–43. doi: 10.1016/j.tics.2008.10.002

Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory*. Hoboken, NJ: Wiley-Interscience.

Davis, M. (2021). The brain-as-computer metaphor. *Front. Comput. Sci*. 3:41. doi: 10.3389/fcomp.2021.681416

De Duve, C. (2003). *Live Evolving: Molecules, Mind, and Meaning*. Oxford: Oxford University Press.

DeCanio, S. J., and Watkins, W. E. (1998). Information processing and organizational structure. *J. Econ. Behav. Organ*. 36, 275–294. doi: 10.1016/S0167-2681(98)00096-1

Dennett, D. C. (1989). *The Intentional Stance*. Cambridge, MA: MIT Press. doi: 10.1017/S0140525X00058611

Downing, K. L. (2015). *Intelligence Emerging: Adaptivity and Search in Evolving Neural Systems*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9898.001.0001

Dunbar, R. I. M. (1993). Coevolution of neocortical size, group size and language in humans. *Behav. Brain Sci*. 16, 681–735. doi: 10.1017/S0140525X00032325

Dunbar, R. I. M. (2003). The social brain: mind, language and society in evolutionary perspective. *Ann. Rev. Anthrop*. 32, 163–181. doi: 10.1146/annurev.anthro.32.061002.093158

Epstein, R. (2016). The empty brain. Aeon.

Equihua, M., Espinosa Aldama, M., Gershenson, C., López-Corona, O., Munguía, M., Pérez-Maqueo, O., and Ramírez-Carrillo, E. (2020). Ecosystem antifragility: beyond integrity and resilience. *PeerJ* 8:e8533. doi: 10.7717/peerj.8533

Escalona-Morán, M., Paredes, G., and Cosenza, M. G. (2012). Complexity, information transfer and collective behavior in chaotic dynamical networks. *Int. J. Appl. Math. Stat*. 26, 58–66. Available online at: https://arxiv.org/abs/1010.4810

Farnsworth, K. D., Ellis, G. F. R., and Jaeger, L. (2017). "Living through downward causation: from molecules to ecosystems," in *From Matter to Life: Information and Causality*, eds S. I. Walker, P. C. W. Davies, and G. F. R. Ellis (Cambridge, UK: Cambridge University Press), 303–333.

Farnsworth, K. D., Nelson, J., and Gershenson, C. (2013). Living is information processing: from molecules to global systems. *Acta Biotheor*. 61, 203–222. doi: 10.1007/s10441-013-9179-3

Fernández, N., Aguilar, J., Pi na-García, C. A., and Gershenson, C. (2017). Complexity of lakes in a latitudinal gradient. *Ecol. Complex*. 31, 1–20. doi: 10.1016/j.ecocom.2017.02.002

Fernández, N., Maldonado, C., and Gershenson, C. (2014). "Information measures of complexity, emergence, self-organization, homeostasis, and autopoiesis," in *Guided Self-Organization: Inception, Vol. 9 of Emergence, Complexity and Computation*, ed M. Prokopenko (Berlin; Heidelberg: Springer), 19–51. doi: 10.1007/978-3-642-53734-9_2

Flack, J. C. (2017). Coarse-graining as a downward causation mechanism. *Philos. Trans. R. Soc. A* 375:20160338. doi: 10.1098/rsta.2016.0338

Fodor, J. A., and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition* 28, 3–71. doi: 10.1016/0010-0277(88)90031-5

Froese, T., and Stewart, J. (2010). Life after ashby: ultrastability and the autopoietic foundations of biological autonomy. *Cybern. Hum. Know*. 17, 7–50. doi: 10.1007/s10699-010-9222-7

Froese, T., and Ziemke, T. (2009). Enactive artificial intelligence: investigating the systemic organization of life and mind. *Artif. Intell*. 173, 366–500. doi: 10.1016/j.artint.2008.12.001

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press; Bradford Books. doi: 10.7551/mitpress/2076.001.0001

Garnier, S., Gautrais, J., and Theraulaz, G. (2007). The biological principles of swarm intelligence. *Swarm Intell*. 1, 3–31. doi: 10.1007/s11721-007-0004-y

Gell-Mann, M., and Lloyd, S. (1996). Information measures, effective complexity, and total information. *Complexity* 2, 44–52. doi: 10.1002/(SICI)1099-0526(199609/10)2:1<44::AID-CPLX10>3.0.CO;2-X

Gershenson, C. (2002). *Contextuality: A Philosophical Paradigm, With Applications to Philosophy of Cognitive Science*. POCS Essay, COGS, University of Sussex.

Gershenson, C. (2004). Cognitive paradigms: which one is the best? *Cogn. Syst. Res*. 5, 135–156. doi: 10.1016/j.cogsys.2003.10.002

Gershenson, C. (2007). *Design and Control of Self-organizing Systems*. Mexico: CopIt Arxives. Available online at: http://tinyurl.com/DCSOS2007

Gershenson, C. (2010). Computing networks: a general framework to contrast neural and swarm cognitions. *Paladyn J. Behav. Robot*. 1, 147–153. doi: 10.2478/s13230-010-0015-z

Gershenson, C. (2011a). *Are Minds Computable*? Technical Report 2011.08, Centro de Ciencias de la Complejidad. *https://arxiv.org/abs/1110.3002*

Gershenson, C. (2011b). The sigma profile: a formal tool to study organization and its evolution at multiple scales. *Complexity* 16, 37–44. doi: 10.1002/cplx.20350

Gershenson, C. (2012). "The world as evolving information," in *Unifying Themes in Complex Systems, Vol. VII*, eds A. Minai, D. Braha, and Y. Bar-Yam (Berlin; Heidelberg: Springer), 100–115. doi: 10.1007/978-3-642-18003-3_10

Gershenson, C. (2013a). The implications of interactions for science and philosophy. *Found. Sci*. 18, 781–790. doi: 10.1007/s10699-012-9305-8

Gershenson, C. (2013b). Living in living cities. *Artif. Life* 19, 401–420. doi: 10.1162/ARTL_a_00112

Gershenson, C. (2015). Requisite variety, autopoiesis, and self-organization. *Kybernetes* 44, 866–873. doi: 10.1108/K-01-2015-0001

Gershenson, C. (2020). "Information in science and Buddhist philosophy: towards a non-materialistic worldview," in *Vajrayana Buddhism in Russia: Topical Issues of History and Sociocultural Analytics*, eds A. M. Alekseyev-Apraksin and V. M. Dronova (Moscow: Almazny Put), 210–218.

Gershenson, C. (2021a). Emergence in artificial life. *arXiv:2105.03216*.

Gershenson, C. (2021b). "On the scales of selves: information, life, and buddhist philosophy," in *ALIFE 2021: The 2021 Conference on Artificial Life*, eds J. Čejková, S. Holler, L. Soros, and O. Witkowski (Prague: MIT Press), 2. doi: 10.1162/isal_a_00402

Gershenson, C., Trianni, V., Werfel, J., and Sayama, H. (2020). Self-organization and artificial life. *Artif. Life* 26, 391–408. doi: 10.1162/artl_a_00324

Gleick, J. (2011). *The Information: A History, A Theory, A Flood*. New York, NY: Pantheon.

Haken, H. (1988). *Information and Self-organization: A Macroscopic Approach to Complex Systems*. Berlin: Springer-Verlag. doi: 10.1007/978-3-662-07893-8

Haken, H., and Portugali, J. (2015). *Information Adaptation: The Interplay Between Shannon Information and Semantic Information in Cognition, Volume XII of SpringerBriefs in Complexity*. Cham; Heidelberg; New York, NY; Dordrecht; London: Springer. doi: 10.1007/978-3-319-11170-4

Harvey, I. (2019). Neurath's boat and the sally-anne test: life, cognition, matter and stuff. *Adapt. Behav*. 1059712319856882. doi: 10.1177/1059712319856882

Heylighen, F. (1999). Collective intelligence and its implementation on the web. *Comput. Math. Theory Organ*. 5, 253–280. doi: 10.1023/A:1009690407292

Heylighen, F., Cilliers, P., and Gershenson, C. (2007). "Complexity and philosophy," in *Complexity, Science and Society*, eds J. Bogg and R. Geyer (Oxford: Radcliffe Publishing), 117–134.

Hidalgo, C. A. (2015). *Why Information Grows: The Evolution of Order, From Atoms to Economies*. New York, NY: Basic Books.

Hölldobler, B., and Wilson, E. O. (2008). *The Superorganism: The Beauty, Elegance, and Strangeness of Insect Societies*. New York, NY: W. W. Norton & Company.

Hopfield, J. J. (1994). Physics, computation, and why biology looks so different. *J. Theor. Biol*. 171, 53–60. doi: 10.1006/jtbi.1994.1211

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/1881.001.0001

Kiverstein, J., and Clark, A. (2009). Introduction: mind embodied, embedded, enacted: one church or many? *Topoi* 28, 1–7. doi: 10.1007/s11245-008-9041-4

Krakauer, D., Bertschinger, N., Olbrich, E., Flack, J. C., and Ay, N. (2020). The information theory of individuality. *Theory Biosci*. 139, 209–223. doi: 10.1007/s12064-020-00313-7

Langton, C. G. (1990). Computation at the edge of chaos: phase transitions and emergent computation. *Phys. D* 42, 12–37. doi: 10.1016/0167-2789(90)90064-V

Lehn, J.-M. (1990). Perspectives in supramolecular chemistry–from molecular recognition towards molecular information processing and self-organization. *Angew. Chem. Int. Edn. Engl*. 29, 1304–1319. doi: 10.1002/anie.199013041

Levin, M. and Dennett, D. C. (2020). Cognition all the way down. Aeon.

Lloyd, S. (2001). *Measures of Complexity: A Non-Exhaustive List*. Department of Mechanical Engineering, Massachusetts Institute of Technology.

Lopez-Ruiz, R., Mancini, H. L., and Calbet, X. (1995). A statistical measure of complexity. *Phys. Lett. A* 209, 321–326. doi: 10.1016/0375-9601(95)00867-5

Malone, T. W., and Bernstein, M. S., editors (2015). *Handbook of Collective Intelligence*. Cambridge, MA: MIT Press.

Marshall, J. A., Bogacz, R., Dornhaus, A., Planqué, R., Kovacs, T., and Franks, N. R. (2009). On optimal decision-making in brains and social insect colonies. *J. R. Soc. Interface*. 6:1065–1074. doi: 10.1098/rsif.2008.0511

Martin, C., and Reggia, J. (2010). Self-assembly of neural networks viewed as swarm intelligence. *Swarm Intell*. 4, 1–36. doi: 10.1007/s11721-009-0035-7

Maturana, H., and Varela, F. (1980). *Autopoiesis and Cognition: The Realization of Living*. Dordrecht: Reidel Publishing Company. doi: 10.1007/978-94-009-8947-4

Mayer-Kress, G., and Barczys, C. (1995). The global brain as an emergent structure from the worldwide computing network, and its implications for modeling. *Inform. Soc*. 11, 1–27. doi: 10.1080/01972243.1995.9960177

Michel, M., Beck, D., Block, N., Blumenfeld, H., Brown, R., Carmel, D., et al. (2019). Opportunities and challenges for a maturing science of consciousness. *Nat. Hum. Behav*. 3, 104–107. doi: 10.1038/s41562-019-0531-8

Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. London, UK: Penguin.

Murcio, R., Morphet, R., Gershenson, C., and Batty, M. (2015). Urban transfer entropy across scales. *PLoS ONE* 10:e0133780. doi: 10.1371/journal.pone.0133780

Packard, N. H. (1988). "Adaptation toward the edge of chaos," in *Dynamic Patterns in Complex Systems*, eds J. A. S. Kelso, A. J. Mandell, and M. F. Shlesinger (Singapore: World Scientific), 293–301.

Passino, K. M., Seeley, T. D., and Visscher, P. K. (2008). Swarm cognition in honey bees. *Behav. Ecol. Sociobiol*. 62, 401–414. doi: 10.1007/s00265-007-0468-1

Prokopenko, M., Boschetti, F., and Ryan, A. J. (2009). An information-theoretic primer on complexity, self-organisation and emergence. *Complexity* 15, 11–28. doi: 10.1002/cplx.20249

Prokopenko, M., Lizier, J. T., Obst, O., and Wang, X. R. (2011). Relating fisher information to order parameters. *Phys. Rev. E* 84:041116. doi: 10.1103/PhysRevE.84.041116

Ramón y Cajal, S. (1899). *Textura del Sistema Nervioso del Hombre y de los Vertebrados: Estudios Sobre el Plan Estructural y Composición Histológica de los Centros Nerviosos Adicionados de Consideraciones Fisiológicas Fundadas en los Nuevos Descubrimientos, Vol. 1*. Madrid: Moya.

Reséndiz-Benhumea, G. M., Sangati, E., Sangati, F., Keshmiri, S., and Froese, T. (2021). Shrunken social brains? A minimal model of the role of social interaction in neural complexity. *Front. Neurorobot*. 15:72. doi: 10.3389/fnbot.2021.634085

Reznikova, Z. (2007). *Animal Intelligence From Individual to Social Cognition*. Cambridge, UK: Cambridge University Press.

Roederer, J. G. (2005). *Information and its Role in Nature*. Heidelberg: Springer-Verlag. doi: 10.1007/3-540-27698-X

Roli, A., Villani, M., Filisetti, A., and Serra, R. (2018). Dynamical criticality: overview and open questions. *J. Syst. Sci. Complex*. 31, 647–663. doi: 10.1007/s11424-017-6117-5

Schaller, R. (1997). Moore's law: past, present and future. *IEEE Spectr*. 34, 52–59. doi: 10.1109/6.591665

Scharf, C. (2021). *The Ascent of Information: Books, Bits, Genes, Machines, and Life's Unending Algorithm*. New York, NY: Riverhead Books.

Searle, J. R. (1980). Minds, brains, and programs. *Behav. Brain Sci*. 3, 417–424. doi: 10.1017/S0140525X00005756

Shalf, J. (2020). The future of computing beyond Moore's law. *Philos. Trans. R. Soc. A Math* 378:20190061. doi: 10.1098/rsta.2019.0061

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Techn. J*. 27, 379–423; 623–656. doi: 10.1002/j.1538-7305.1948.tb00917.x

Smolensky, P. (1988). On the proper treatment of connectionism. *Behav. Brain Sci*. 11, 1–23. doi: 10.1017/S0140525X00052432

Solé, R., Amor, D. R., Duran-Nebreda, S., Conde-Pueyo, N., Carbonell-Ballestero, M., and Monta nez, R. (2016). Synthetic collective intelligence. *Biosystems* 148, 47–61. doi: 10.1016/j.biosystems.2016.01.002

Steels, L., and Brooks, R. (1995). *The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents*. New York City, NY: Lawrence Erlbaum Associates.

Stewart, J. (1995). Cognition = life : Implications for higher-level cognition. *Behav. Process*. 35, 311–326. doi: 10.1016/0376-6357(95)00046-1

Stewart, J., Gapenne, O., and Di Paolo, E. A. (eds.). (2010). *Enaction: Toward a New Paradigm for Cognitive Science*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262014601.001.0001

Trianni, V., and Tuci, E. (2009). "Swarm cognition and artificial life," in *Advances in Artificial Life. Proceedings of the 10th European Conference on Artificial Life (ECAL 2009)*. Hungary.

Turing, A. M. (1937). On computable numbers, with an application to the entscheidungs problem. *Proc. Lond. Math. Soc*. s2–42, 230–265. doi: 10.1112/plms/s2-42.1.230

Turing, A. M. (1950). Computing machinery and intelligence. *Mind* 59, 433–460. doi: 10.1093/mind/LIX.236.433

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/6730.001.0001

von Baeyer, H. C. (2005). *Information: The New Language of Science*. Cambridge, MA: Harvard University Press.

von Neumann, J. (1958). *The Computer and the Brain*. New Haven, CT: Yale University Press.

Walker, S. I. (2014). Top-down causation and the rise of information in the emergence of life. *Information* 5, 424–439. doi: 10.3390/info5030424

Wheeler, J. A. (1990). "Chapter 19: Information, physics, quantum: the search for links," in *Complexity, Entropy, and the Physics of Information, volume VIII of Santa Fe Institute Studies in the Sciences of Complexity*, ed W. H. Zurek (Reading, MA: Perseus Books), 309–336.

Wittgenstein, L. (1999). *Philosophical Investigations, 3rd Edn*. Upper Saddle River, NJ: Prentice Hall.

Wolfram, S. (2002). *A New Kind of Science*. Champaign, IL: Wolfram Media.

Zarkadakis, G. (2015). *In Our Own Image: Savior or Destroyer? The History and Future of Artificial Intelligence*. Pegasus Books.

Zubillaga, D., Cruz, G., Aguilar, L. D., Zapotécatl, J., Fernández, N., Aguilar, J., et al. (2014). Measuring the complexity of self-organizing traffic lights. *Entropy* 16, 2384–2407. doi: 10.3390/e16052384

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Computer Science

TYPE Hypothesis and Theory
PUBLISHED 09 January 2023
DOI 10.3389/fcomp.2022.976801

Nine insights from internet engineering that help us understand brain network communication

Daniel J. Graham*

Department of Psychological Science, Hobart and William Smith Colleges, Geneva, NY, United States

Philosophers have long recognized the value of metaphor as a tool that opens new avenues of investigation. By seeing brains as having the goal of representation, the computer metaphor in its various guises has helped systems neuroscience approach a wide array of neuronal behaviors at small and large scales. Here I advocate a complementary metaphor, the internet. Adopting this metaphor shifts our focus from computing to communication, and from seeing neuronal signals as localized representational elements to seeing neuronal signals as traveling messages. In doing so, we can take advantage of a comparison with the internet's robust and efficient routing strategies to understand how the brain might meet the challenges of network communication. I lay out nine engineering strategies that help the internet solve routing challenges similar to those faced by brain networks. The internet metaphor helps us by reframing neuronal activity across the brain as, in part, a manifestation of routing, which may, in different parts of the system, resemble the internet more, less, or not at all. I describe suggestive evidence consistent with the brain's use of internet-like routing strategies and conclude that, even if empirical data do not directly implicate internet-like routing, the metaphor is valuable as a reference point for those investigating the difficult problem of network communication in the brain and in particular the problem of routing.

KEYWORDS

computer metaphor, internet metaphor, systems neuroscience, theoretical neuroscience, packet switching, brain dynamics, network communication, computational neuroscience

Check for updates

OPEN ACCESS

EDITED BY
Pedro Martinez,
University of Barcelona, Spain

REVIEWED BY
John Mollon,
Cambridge University, United Kingdom
Saket Navlakha,
Cold Spring Harbor Laboratory,
United States

*CORRESPONDENCE
Daniel J. Graham
✉ graham@hws.edu

SPECIALTY SECTION
This article was submitted to Theoretical Computer Science, a section of the journal Frontiers in Computer Science

RECEIVED 05 July 2022
ACCEPTED 12 December 2022
PUBLISHED 09 January 2023

CITATION
Graham DJ (2023) Nine insights from internet engineering that help us understand brain network communication.
Front. Comput. Sci. 4:976801.
doi: 10.3389/fcomp.2022.976801

> *Metaphor consists in giving the thing a name which belongs to something else.*
> Aristotle, *Poetics* xxi, tr. Bywater
> *Mathematics is the art of giving the same name to different things.*
> Henri Poincaré, *The Future of Mathematics*, 1908

Introduction

Philosophers have long recognized that the development of a new metaphor can encourage researchers to take unorthodox ideas seriously (Bartha, 2022). In the sciences, new metaphor can spur theorists to build classes of models different from those that already exist. Each new metaphor succeeds not by capturing the exact workings of the analogized system but rather by giving us a new vision of some otherwise unapproachable entity. Theory in the physical sciences has been especially reliant on insights from a succession of metaphors, each an improvement on its predecessor: the container space metaphor for the physical universe gives way to Einstein's fabric of space time.

Metaphor is just as important if not more so to biological theory. Its foundational idea, Darwinian evolution, was crystalized in the metaphor of a tree. Darwin's tree of life was not literally a tree—all life does not spring forth from a single plant. Instead, the metaphor brings together several key properties of the system: rootedness, or the idea that the base of the system of living organisms on earth has one or a small number of main roots; divergence, or the idea that branches spread out and bifurcate, but rarely inosculate (rejoin); and relatedness, or the historical dependence and elaboration of distal twigs on proximal branches. Though graphical depictions of various proposals for the chain of life preceded Darwin, no one before him had seen the problem in this way. The metaphor has proven transformative. It remains in common use today even as knowledge of phylogenetic complexity unknown to Darwin has accumulated (Quammen, 2018).

We needed a metaphor for the brain, and "the computer" has served us well

As attested by the present Research Topic articles—and indeed most issues of any research journal in the neurosciences—researchers rely on the computer metaphor when studying the brain, even if they disagree about its formulation and in what way it is useful (e.g., Richards and Lillicrap, 2022). Historically, McCulloch and Pfeiffer (1949) saw single neurons as a transistor in a "multi-gridded" brain. Most prominently today, the metaphor inheres when neuronal "representation" is seen as having the effect of generating elements of Turing machine symbols and operations (Richards and Lillicrap, 2022), or when neuronal tuning properties are seen to serve as elements in a particular code (e.g., Olshausen and Field, 1996). One thing different instantiations of the computer metaphor seem to have in common is that they see things from the point of view *representational elements* (see also Poldrack, 2021; Anderson and Champion, 2022; Brette, 2022; Hipólito, 2022; John, 2022). In this view, activity in a given neuron embodies an act of representation in one form or another (see e.g., Baker et al., 2022). A given pattern of activity in neurons and/or across neuronal populations is seen to indicate the brain's invocation of a particular coding element (e.g., for visual data, as a basis function, or as the features in some layer of a convolutional neural network).

The "brain-as-representation machine" metaphor is also made visible in works such as Gidon et al. (2022). These authors propose a thought experiment regarding the nature of consciousness and ask whether "replay" of neuronal signals via external means is equivalent to an identical endogenous experience. Whatever one thinks about the thought experiment, it assumes brain function consists only of representational processes, to the point where the authors illustrate the procedure of the thought experiment with cartoon "play" and "record" icons.

This view concretizes a particular understanding of the brain's goals, and facilitates the importation of ready insights, tools, and methods from other fields, especially mathematics, to attack the difficult problem of understanding the purpose and meaning of neuronal signals. This effort has propelled the field through a period of rapid advances in the 20th and 21st centuries (Cobb, 2020; Lindsay, 2021). The metaphor helped identify a problem to be solved, and offered a range of more and less literal implementations to consider. Even as the limits of the metaphor are probed, it retains value as an impetus and sometimes a foundation for more precise understanding.

A new metaphor: The internet

As useful as the representational metaphor is, it cannot capture all system goals when the system is as complex as the brain. Brains instantiate many goals. For this reason, not all signals extracted from the brain should necessarily be seen to serve the goal of representation. Here I argue for another class of metaphors that we can invoke in addition to other metaphors: the internet (Danilova and Mollon, 2003; Graham and Rockmore, 2011; Oka et al., 2015; Graham, 2021).

The internet and the brain are clearly different, just as physical computers, and indeed Turing machines, are different from the brain. But, taking inspiration from the history of computational neuroscience and its metaphorical framework, we can profit from considering the conceptual infrastructure for communication on the internet as a point of reference to the problem space of network communication in the brain.

This can help us determine which neuronal signals relate to communication and which to representation, and in what way representation and communication relate to each other.

If neurons compute, there is of course a superficial correspondence between the brain as a whole and the internet, since both systems involve the networked linkage of many localized computational units. But one can't simply wire computers together and expect them to communicate reliably. Even the simplest computer networks of the Web 1.0 era required "phenomenally complex" network engineering (Meyers, 2004). A comprehensive and cohesive conceptual framework is needed to make it work.

In adopting the internet metaphor, we attempt to see the brain from the point of view of *messages*, rather than representational elements. In neuronal terms, this shift implies a consideration not only of how neurons relate to environmental inputs and behavioral outputs—"outside-in neuroscience"—but especially a consideration of how neurons relate to each other—an "inside out" approach (Buzsáki, 2019; Fields et al., 2022; Mayner et al., 2022). More specifically, the goal is to understand how the brain's vast and interconnected network of elements organizes message passing within itself by examining a variety of possible schemes for communication (Graham et al., 2020). This approach is consonant with other integrative conceptions of brain function such as neural re-use (Steriade, 2004; Anderson, 2010), neuronal recycling (Dehaene, 2005), computational flexibility (Pessoa et al., 2019), and emergence (e.g., Varley and Hoel, 2022), among others, and can be seen as a way to bring these related proposals together under a common and more concrete framework.

Historically, the goal of understanding network communication was an initial impetus for the cybernetics movement and has antecedents going back at least to Spencer (1896). Pavlov (1927) and Sherrington (1947) highlighted the problem as well, in part by making a comparison with telephone and rail networks. But the advent of the modern internet in the second half of the twentieth century, based on the conceptual underpinning of packet-switched networking, transformed understanding of distributed network communication. This development had ramifications far and wide, and brain science soon took notice. Just ahead of the launch of NSFNet, Poggio (1984) had begun to sketch out a fundamental role for routing in the brain, using the existing ARPANET's packet-switched routing system as an analogy. Since that time, others have built models of routing on brain networks, though such ideas do not always explicitly reference the internet. These include the dynamic routing model (Olshausen et al., 1993) and the notion of routing by synchrony (also called communication through coherence: Fries, 2005; Mishra et al., 2006; Nádasdy, 2010), with additional routing-based insights being offered by Wolfrum (2010) and Navlakha et al. (2018), among others. The present work is an attempt to unify and advance these investigations via a more systematic examination of the characteristics of effective routing, and to point out some of the challenges inherent in network communication. Of particular importance is how the internet flexibly deals with interacting signals that make use of shared resources.[1]

What is routing?

Routing systems govern how messages travel among nodes that are connected by links. Internet protocol embodies one routing strategy, while other strategies include those underlying postal and traditional telephone systems. Routing is necessary when communicating nodes are separated in space by distances much larger than the size of a node, and when nodes are not all directly connected to one another. As such, routing requires a degree of mutual trust among nodes and a preparedness for faults and errors.

Though it is implemented locally, routing allows nodes across the network to select different targets across the network at will (Graham, 2014). Routing presumes that some nodes can receive messages over multiple incoming edges and transmit them over multiple outgoing edges, based on some rules or algorithms. In the case of converging inputs, routing rules arbitrate among messages arriving on different incoming edges. When outgoing edges diverge, routing serves to direct messages on outgoing edge(s). Routing thus serves to manage congestion and enable flexibility in message passing.

Routing strategies become irrelevant if the number of incoming and outgoing edges at all nodes is the same and messages arriving at a node on edge *a* always leave on edge *b*. However, the term "node" in this case loses its meaning. Each path of incoming and outgoing edges through the node can be simplified in a network description as a single edge, and the node and indeed the network as such disappears.

Routing processes in the brain

While brains can manage message flow by reorganizing connectivity (e.g., Fauth and Tetzlaff, 2016), this process is too slow to direct neuronal signals over millisecond, second, and minute timescales. Even if changes in network structure do alter message flow, routing processes are still necessary to achieve reliable, selective communication. Thus, if it is at all sensible to describe brains as networks composed of nodes and edges, then we need to consider how to find and execute paths on an essentially fixed network, and how signal interactions might be managed in brains like ours that have no central controller.

[1] Others have noted additional metaphorical links to how applications on the internet, such as the World Wide Web, organize distributed information (Varela et al., 2001; Griffiths et al., 2007). However, these applications do not relate to routing per se, which is the focus of the current paper.

What evidence is there that brains need to perform routing of the kind described here? Though direct measurement of signal flow over structurally identified neuronal networks is not yet robustly achievable, there are many levels of organization where the need for routing is apparent, and where suggestive evidence of routing processes has been found.

In terms of brain region structural connectivity, there is clearly the possibility of routing, even if it is not normally described in this way. Treating regions as nodes, with messages incoming and outgoing along white matter tracts, signals arising in a given region—say V1—can be sent via axons originating in layer 2/3 or 4 directly to other regions (say V2), or along projections via the thalamus to other cortical regions (with potential for modulation of these signals by the cortical target), or to thalamus and back to a different part of V1, or to other cortical areas, and then on to propagate to other destinations (see e.g., Reichova and Sherman, 2004; Anderson and Martin, 2016). Routing in this core of the network can happen very quickly: signals can be relayed on round-trips between thalamus and cortex in as little as 9 milliseconds (Briggs and Usrey, 2007). Though white matter signals may arrive in structurally segregated parts of a region, they stand a good chance of interaction given the high interconnectivity within regions, and therefore appear subject to some system of routing.

Brain imaging studies have given functional indications of routing at the regional level. In humans, Cole et al. (2013), found evidence that frontal and parietal areas flexibly communicate with different modalities as well as other systems (e.g., motor) at different times. Gerraty et al. (2018) found evidence that striatal nuclei can selectively engage different cortical targets in different behavioral contexts. Mechanisms that allow this kind of routing may involve synchronization of subthreshold oscillations between or among areas (Singer, 1999; Fries, 2005; Womelsdorf et al., 2007; Nádasdy, 2010; Gisiger and Boukadoum, 2011; Palmigiano et al., 2017; Javadzadeh and Hofer, 2021; Boroujeni and Womelsdorf, 2022; Sakalar et al., 2022). Oscillatory mechanisms may also contribute to routing functions within regions (e.g., communication between subpopulations in V1; Gray et al., 1989).

At the level of single neurons, the ability to route or "steer" messages on different paths has long been posited for single neurons (Waxman, 1972; Scott, 1977). Several single neuron-level cortical mechanisms have recently been observed that could dynamically manage incoming messages. For example, input selection may be partly shaped by exclusive-or (XOR) gating at dendrites (Gidon et al., 2020) or via other dendritic gating mechanisms (Steriade and Paré, 2007; Gollisch and Meister, 2010; Oz et al., 2021). Steering via axon gating is also possible given considerable axonal branching in cortex (Winnubst et al., 2019) which has long been suspected to allow transmission control at branch points. Axonal mechanisms of routing could also involve axon-axon interactions (e.g., Epsztein et al., 2010). Other mechanisms that could perform routing at the single neuron level have been suggested such as ephaptic interactions (Sheheitli and Jirsa, 2020); glia-mediated synapses (Möller et al., 2007); and local spreading of neuroendocrine molecules (Bargmann and Marder, 2013). Probabilistic modeling of signal transmission among four neurons in hippocampus provides suggestive evidence consistent of a highly flexible capacity for routing in the brain (Nádasdy et al., 1999), though this study's results can be interpreted in other ways; see Section Introduction and Box 1 for a detailed discussion of this study.

To integrate and understand how these kinds of neurobiological mechanisms may be deployed to perform routing, it is helpful to consider the strategies and goals that led to the construction of the modern internet. I offer nine insights that helped make the modern internet possible and begin to apply these ideas to the brain. The goal is to move toward more concrete models and hypotheses, though these are yet to be developed.

Nine insights from internet engineering applied to network communication in the brain

Internet routing protocol specifies communication procedures and standards across essentially all modern computer and mobile device networks. However, the internet is defined not so much by its physical implementation in linkages among devices but rather as a set of rules governing the treatment of messages. The core framework for internet communications is the open systems interconnection (OSI) model. The OSI model is not a theory rooted in basic mathematics or physics. Rather, it comprises two broad branches: (1) a conceptual architecture for overall engineering design to route messages successfully and (2) a hierarchy of protocol standards. The simplified "layers" of the OSI protocol model are briefly summarized in Figure 1.

The design goals of the OSI model and their implementation on the modern internet are especially relevant for the study of processes of routing in the brain, or what might be called "communicatory neuroscience." The following sections highlight strands of neurobiological evidence that are suggestive of—but do not verify—neuronal implementations of sophisticated routing strategies in the mammal brain; two lines of evidence are examined in more detail and framed as interrogatives in Boxes 1, 2.

Insight 1: Routing must be flexible

The internet's flexibility is its central goal. In terms of function, any sender and receiver, no matter their degree of separation on the vast network, can communicate at will

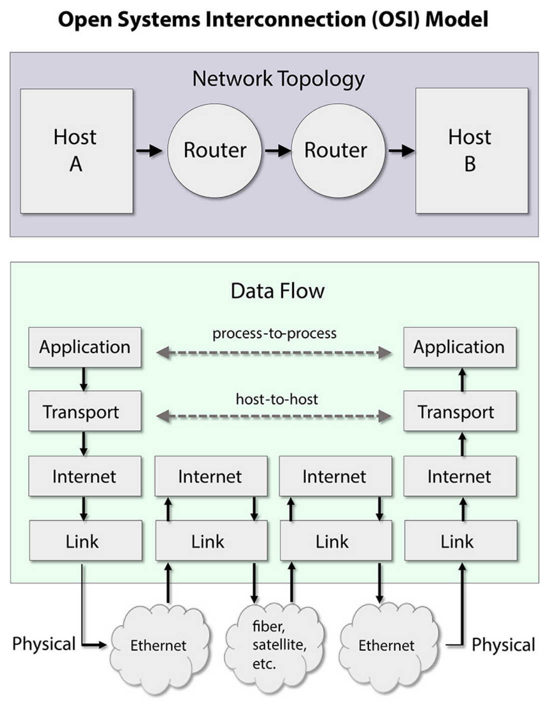


FIGURE 1
A simplified conceptual design of the internet protocol stack, based on the open systems interconnection (OSI) model. Network topology determines the passage of messages across the network between hosts. The flow of data across the network is organized into conceptual layers. A message originates in the application layer and descends by way of the transport layer and internet layer to a physical layer link (e.g., wire or fiber-optic cable). At intermediary routers, messages ascend only to the internet layer, which plans out the message's forward route. The message then returns to the physical layer for onward travel. This diagram omits the presentation layer and the session layer, which are less relevant for our purposes, and can be seen as part of the application layer.

with each other, as long as a limited set of protocol is followed. Crucially, flexible communication is delivered over shared resources. As Danilova and Mollon (2003) observe, "The essential feature of the Internet is that it eliminates the need for a dedicated cable between any particular pair of computers that need to communicate." Flexibility is needed not only in who communicates with whom, but also what path messages take once targets are chosen (see Insight 2) and what kinds of information nodes exchange (Insight 5). Achieving the overarching goal of flexibility shapes all other features, and these features are described in the remaining sections.

Flexibility is similarly fundamental to the brain. Full interconnectivity is impossible: in the human brain, it would require a 20 km-wide head (Nelson and Bower, 1990). Moreover, the behaviors and tasks brains need to accomplish in the world and the brain's network infrastructure strongly suggest flexible control of information flow (Kreiter, 2020; Safron et al., 2022). The most well-developed models of neuronal mechanisms for this kind of flexibility relate to perceptual invariances (Olshausen et al., 1993; see also Wiskott, 2006) and attention (e.g., Mishra et al., 2006) but other "outside-in" functions like flavor perception, decision making, reasoning, problem-solving, sociality, planning, language, creativity, and many others also plainly require flexible management of information flow. The brain's routing strategies must support the accomplishment of highly varied tasks based on highly varied inputs, and do so on a network structure that is fixed in the short term. To take one example, it is possible for the brain to extract different information from the same scene or context depending on one's goal (Günseli and Aly, 2020). Likewise, decision making, whether modeled as evidence accumulation in frontal neurons (Gold and Shadlen, 2007), or as some other "choosing" process, must include delivery of chosen outputs to distinct neuronal subsystems along paths that were equally viable before the decision was "made." Flexibility may also help the brain to reroute signals around focal lesions without growing new connections (Zalesky et al., 2007; Fornito et al., 2015).

From an "inside-out" perspective, flexibility is suggested by the evidence noted above of selective targeting at the regional level and by the fact that there are numerous short paths between most pairs of regions (considered further in the next section). At the single cell level, suggestive but not conclusive evidence for flexible steering of signals comes from the *in vivo* electrophysiological study of Nádasdy et al. (1999) discussed in Box 1. However, tracing signal propagation across neuronal networks with known connectivity, which could provide more conclusive evidence of flexible routing, remains an unsolved problem in neuroscience (see Box 1).

Insight 2: Routing should take advantage of network structure

The founders of the modern internet saw that network topology and the design of routing protocol were inextricably linked. The two key innovations that led to the internet—distributed network architecture and packet-based protocol—were conceived in tandem by Baran (1964); see also Boehm and Baran (1964).[2] I will deal with the effects of network architecture first, and consider its packeted nature in the next section.

Baran realized that a distributed network—one that compromised between a star-shaped network and a lattice—would allow short paths between almost any pair of nodes,

[2] British researcher Donald Davies made essentially the same two proposals, also in 1964.

BOX 1 Evidence for flexible routing in hippocampal circuits?

In Nádasdy et al. (1999), extracellular tetrode recordings were obtained from hippocampal CA1 pyramidal layer neurons in 18 rats during sleep and conditioned wheel running. Clustering was performed on multi-channel signals to identify four individual neurons. The researchers then used Monte Carlo models to track temporal patterns of spikes as they appeared to propagate between four hippocampal neurons. In particular, they used shuffling of spike train patterns to identify patterns of spike timing in different neurons that could not be reasonably explained as chance occurrences. They interpreted these spike trains as messages passed from neuron to neuron, which allows one to trace their putative paths of propagation, bearing in mind that ground-truth connectivity was not measured. Some of the results obtained from this analysis are shown in Figure 2.

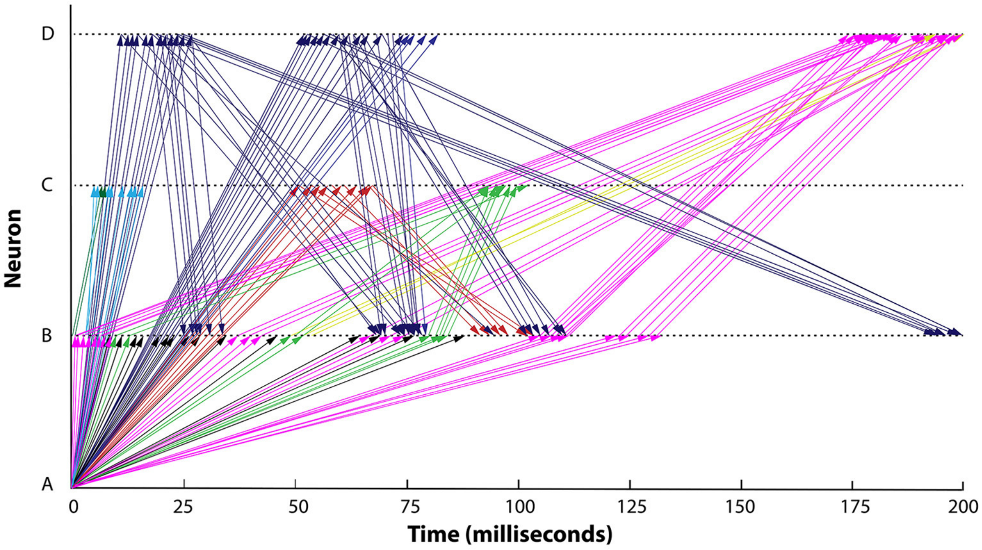


FIGURE 2
Exchange of spike train messages among four neurons over time in the rat hippocampus during wheel running. Messages are seen to pass among four neurons, labeled A–D. Colors indicate messages traveling on the same path. Horizontal axis indicates time (0–200 ms). Data from Nádasdy et al. (1999), figure redrawn from Buzsáki (2004).

Interpreting these signal transmissions as messages, as the authors do, single neurons appear to have the ability to direct signals on different paths to the same or different targets. These patterns of message flow also vary systematically between behavioral conditions (see Figure 4 in Nádasdy et al., 1999). Arrival times of spike train messages show both short and long delays (latencies), indicating that messages may travel over one or more intermediaries when traveling between the measured neurons (some delays were over 100 milliseconds). It is important to note that each of the message paths suggested here is not a one-off, but is rather a path observed at least a dozen times, which suggests that polysynaptic (multi-hop) transmission is reliable. In sum, one interpretation of these data is that the routing protocol that controls this subnetwork allows all of the following flexible behaviors:

- Sending messages to different destinations.
- Sending messages on different paths to the same destination.
- Sending messages on a given route with a small or large variation in timing.
- All of the above on polysynaptic paths.
- Flexibly changing routing in different behavioral contexts.

As noted above, connectivity was not measured in this study. One certainly cannot rule out the possibility that the observed patterns are artifacts of analysis, or epiphenomena. It could certainly be the case that cells not recorded from are driving the four cells studied. For example, a given "control" cell could produce a particular spiking pattern, which could be relayed by four sets of intermediaries that provide different delays such that that pattern appears at its observed targets at corresponding times. Yet this interpretation would not necessarily invalidate the view that the system is demonstrating flexible routing. Intermediaries would need to faithfully transmit the control message of the spike train in mostly unaltered form and they would need to "protect" these messages from interference from other incoming signals, all while being able to change control patterns reliably both within and across behavioral states. For a common control cell to generate different patterns of delay in the four observed neurons, the intermediaries would need to dynamically change their latencies, and/or selectively direct messages on different intermediary paths.

Nevertheless, the results of Nádasdy et al. (1999) are ambiguous. More than two decades after this study, it remains very difficult to trace signals as they traverse multiple nodes of known connectivity in a brain network (see van der Meij and Voytek, 2018; Hodassman et al., 2022). Models that rely on inferring causality linking separate measurements of structure and activation (e.g., Javadzadeh and Hofer, 2021) can be misleading (see, e.g., Mehler and Kording, 2018; Brette, 2019; Bruineberg et al., 2021).

But though tracing signal propagation faces great procedural challenges, part of the reason why studies that directly trace signal propagation remain rare may be that we have not yet fully appreciated the challenges of flexible routing. As a result, we have limited expectations about what neuronal signatures to expect. Often, we see a neuron's "job" as participating in a computation or representation, where correlations between predicted patterns of activity and observed activity in a given context are seen as sufficient evidence that the brain is carrying out the proposed computation. Approaches like Nádasdy et al. (1999), on the other hand, see spike trains as indications that there are messages to be propagated (see also Luczak et al., 2013; discussed in Insight 4; Grosmark and Buzsáki, 2016). In this view, some aspect of a spike-based message passed between neurons maintains coherence as it propagates, though its structure may be subject to new transformations as it travels—analogously to the way an internet data packet is wrapped in different containers at different points in its journey (e.g., frames and flows). Approaches that build on this insight may lead to advances in our understanding.

BOX 2 Could thalamo-cortical loops deliver message acknowledgments (ACKs)?

The thalamus lies near the center of the human brain, and appears to play the role of network backbone (see Hilgetag et al., 2016). Under the "higher-order relays" picture of connections between thalamus and cortex (Sherman and Guillery, 1998, 2001, 2002; see Figure 3), the thalamus contains first-order relays (e.g., lateral geniculate), which receive inputs from the sensorium (e.g., retina). First-order relays pass those inputs on to first-order cortical targets (e.g., V1), which reciprocate back to the same area of the thalamus. The thalamus also contains higher-order relays such as the pulvinar, which receive input from first-order cortical territories, and have connections back to those first-order areas, as well as connections to "higher-order" cortical areas (e.g., V2). In this way, information can travel widely in the cortex in just a few hops via thalamic relays.

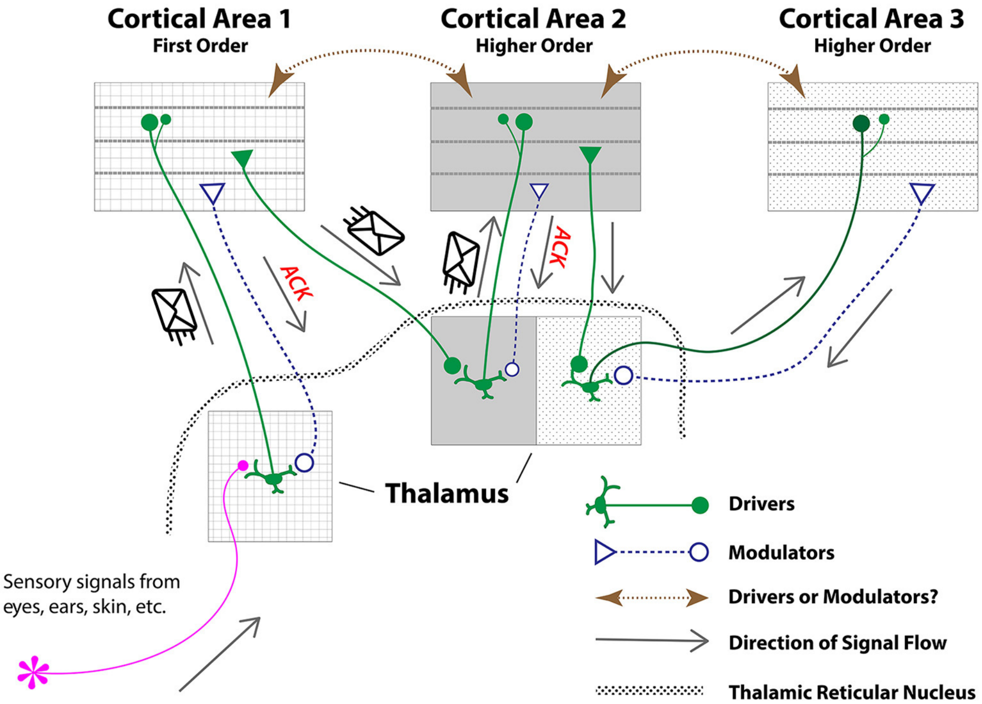


FIGURE 3
A schematic, hypothetical model under which thalamic relays provide ACKs over the "higher-order relays" organization of thalamo-cortical circuits (Sherman and Guillery, 1998, 2001, 2002). Under the scheme proposed in the current paper, messages containing "content" are sent by "driver" neurons, while "modulators" return ACK-like messages back to thalamic senders, either directly or by way of the thalamic reticular nucleus. If the driver's message is delivered successfully, modulator ACK messages would prevent resending. If a timely ACK is not received, a driver in thalamus could be triggered to resend the missing message. Note that an ACK sent from cortex to thalamus confirming receipt implicitly confirms that the message successfully traveled on an earlier leg from cortex to thalamus (since thalamic excitations are seen as signals relayed from elsewhere), perhaps obviating the need for ACKs on the earlier leg. Figure adapted from Reichova and Sherman (2004).

Given the comparatively large distances traveled between cortex and thalamus and the possibility of spike failure (which is more common in long axons), as well as other types of message loss or corruption, some system of delivery verification would seem appropriate for this core of the brain network. The conventional idea that descending connections serve to "adjust the weights" of incoming signals (e.g., as a way to modulate attention) does not explain why long loops to the thalamus would be required—weights could in principle be adjusted by local circuits in cortex itself, without the cost, delay, and risk of making a long projection back to the thalamus. Instead, this architecture appears better suited to flexible and verifiable message passing among cortical areas via the thalamus.

Cortico-thalamo-cortical communication in the Sherman and Guillery picture is thought to be mediated by two parallel links that go in opposite directions: (1) "driver" connections originate in higher-order thalamic nuclei, traveling to higher-order cortical areas (drivers are also considered to include projections from first-order thalamic relays, such as LGN, to first-order cortical areas, such as V1, and projections from layer 5 of first-order cortical areas to higher-order thalamic relays, such as pulvinar); (2) "modulator" projections originate in layer 6 of the areas targeted by higher-order thalamic relay drivers, and descend to the thalamus to synapse onto the dendritic arbors of drivers (Sherman and Guillery, 1998, 2001, 2002).[3] Drivers form a minority of inputs but are seen to deliver primary messages. Modulators are much more numerous, and can affect the likelihood of transmission of driver signals but do not seem to alter the content of those signals (e.g., they do not change receptive field properties of first-order thalamic relays from LGN to V1; Reichova and Sherman, 2004). Modulators also connect to the thalamic reticular nucleus, which can exert inhibitory influence on most connections between the thalamus and cortex.

[3] Signals also travel via far more numerous cortico-cortical connections within gray matter. These connections seem to be classifiable as drivers or modulators (Sherman and Guillery, 2011) and could conceivably support an acknowledgment system that runs in parallel to the postulated thalamocortical system. However, such direct, short-range cortico-cortical connections may be reliable enough to not require ACKs.

(Continued)

BOX 2 (Continued)

By thinking of the brain in terms of messages and routing, we can sketch a scheme by which thalamic relays could provide ACKs. Messages containing "content" are sent by drivers, while modulators return ACK-like messages back to thalamic senders, either directly or by way of the thalamic reticular nucleus. If the message is delivered successfully, modulator ACK messages would prevent resending via inhibition. If a timely ACK is not received, a driver in thalamus could be triggered to resend the missing message. In this scheme, ACKs are not performed to confirm receipt of driver messages sent from cortex to thalamus. This may be a sensible strategy. An ACK sent from cortex to thalamus confirming receipt implicitly confirms that the message successfully traveled on an earlier leg from cortex to thalamus (since thalamic excitations are seen as signals relayed from elsewhere). Too many ACKs can clog a system so providing ACKs on only half of each loop could make better use of bandwidth. One would predict that a capacity exists in the thalamus (possibly in the reticular nucleus) for buffering in case ACKs are not received in the thalamus and driver messages need to be resent.[4]

Matsuyama and Tanaka (2021) have recently found *in vivo* electrophysiological evidence of "switch-type" neurons in higher-order thalamic nuclei in primates that produce strong bursts after initial visual-auditory stimulus presentation (flash and tone), but become suppressive with repetition (see also Guo et al., 2017; Sieveritz and Raghavan, 2021). This kind of behavior could serve as a building block for a system of ACKs like that described here (see also Crabtree, 2018). However, more detailed models than can be offered here are needed. Indeed, the message acknowledgment scheme proposed here is merely a first step toward a model under a reconceptualization of brain networks as communication systems, which co-exist with computational architectures. The scheme is not a model in and of itself. It should also be emphasized that a solution like ACKs might make sense in cortical-subcortical loops but would not make sense in, for example, spinal reflex arcs, where motor fibers need not receive neural feedback from muscles, but can rather rely on sensory feedback directly.

but without central switchboards. Short characteristic path length (i.e., low average shortest path length) would later be recognized as a defining property of "small-world" networks, along with high clustering (Watts and Strogatz, 1998). Routing design can take advantage of networks with short paths between nodes. On the modern internet, this is achieved through backbone nodes and peering, i.e., building short cuts between subnetworks to achieve robust interconnection of diverse entities spread across large distances. Not only are paths short on a distributed network, Baran realized, there are usually multiple short paths available, allowing compensation for lost nodes and links, as well as for changes in traffic volume. The system is designed specifically so that, as new conditions arise, new routes are chosen, even as network structure remains the same. This has been termed "robust yet fragile" behavior (Li et al., 2004; Doyle et al., 2005; Sneppen et al., 2005).

In the brain, the connectomics movement has shown that network architecture is also characterized by short average path lengths between nodes (see, e.g., Sporns, 2012). Cortical areas of the macaque monkey are on average about 1.5 hops from each other, and in the mouse the value is closer to 1 (e.g., Knoblauch et al., 2016; Gămănut et al., 2018). The value for the entire primate connectome is not known but I predict it is around 3 or 4 for most pairs of neurons (see also Parsons et al., 2022). This implies that a given brain component can and does interact with most other components via redundant short paths.

Moreover, connectivity in the brain is redundant at multiple levels. Populations of e.g., neurons tuned to the same feature such as orientation columns, are usually connected to common target populations. At the level of brain regions, the network statistic of "communication efficiency" (Latora and Marchiori, 2001) gauges the number of parallel short paths between a given pair of nodes. This and related measures ["search information" based on the measure of Rosvall et al. (2005)] are found to be accentuated in brain networks, and conducive to effective communication, in comparison to randomly rewired networks of the same degree sequence (see e.g., Avena-Koenigsberger et al., 2017; Seguin et al., 2018).

However, in the brain, the existence of short paths implies that signals passed between components stand a good chance of interacting with each other en route, potentially in deleterious ways. This problem necessitates systematic routing strategies. The likelihood of signal interactions on networks is greatly reduced if the network has a different architecture, such as a lattice or a tree, but this would engender longer paths (but note that some network architectures that differ from that of the internet and that of the connectome, such as random Erdos-Renyi graphs, also have short characteristic path lengths). Shortest path measures are often used in network neuroscience to evaluate the ways that network architecture affects communication among nodes. However, shortest paths are only short if there is no possibility of message interaction, and therefore of errors, congestion, and delay. As Seguin et al. (2019) have argued, it is implausible that the brain has global awareness of network structure necessary for finding all shortest paths (but see Mišić et al., 2015). But it is less plausible still that the brain can always use shortest paths without running into congestion. Instead, in the following sections, I consider local, protocol-based approaches to management of short (but not necessarily shortest) paths and specifically how the internet packages messages and shares communication links. These strategies can serve as potential points of reference for how the brain achieves parsimonious and reliable movement of messages.

[4] On the internet, nodes use buffers to perform queueing, or lining up incoming messages in a small memory allocation based on when they arrived, and directing them on the proper outgoing path one-by-one. In the brain, hypothetical buffers might only need to store a single message, and for only a brief period. Delay circuit-like mechanisms for such buffers have been proposed (Goldman-Rakic, 1996; Funahashi, 2015), and some models of connectome dynamics include node buffers (Mišić et al., 2014; Fukushima and Leibnitz, 2022). Buffers remain hypothetical but with the impetus of the internet metaphor, they invite further investigation.

The internet has additional architectural motifs such as scale-free architecture (Barabási and Albert, 1999; Caldarelli et al., 2000) and rich clubs (Zhou and Mondragón, 2004; Colizza et al., 2006). Brains show some of these motifs (e.g., Van den Heuvel and Sporns, 2011). However, wiring patterns in the brain are diverse. We should expect that specialized motifs will shape the design of the brain's routing strategies. But these motifs should co-exist with global rules and the network-wide phenomenon of short average path lengths. In this context message interactions must be managed. Because they produce nonlinear effects, principled numerical simulations of routing protocols on brain networks may help us uncover novel relationships between network structure and message interactions on networks (see Hao and Graham, 2020).

Insight 3: Routing can exploit shared resources

The existence of routing presupposes that one has specified the nature of a message. Paul Baran's second insight related to the structure of messages. He realized that message components need not be sent in contiguous units of arbitrary size, the way a phone conversation or a postal letter is. Instead, messages can be divided up into equal-sized chunks—packets—and spread through the network dynamically. This approach was married to a strategy of sharing resources and treating everyone's packets as interchangeable. Sharing in this way requires a leap of faith that "my" message parts won't get lost among those belonging to everyone else as they travel across the network, since no one has exclusive access to intermediary links. Baran and others deliberately imbued each part of the network with sharing and with trust in the wider network—this is the "openness" of the OSI model. Organizing the use of shared resources over an open, small-world network is accomplished by a collection of communication engineering tricks, which are described in the remaining sections.

If communication resources in the brain are shared, as connectome structure described in Insight 2 implies, the system might need to employ solutions like those of the internet. It is worth considering if a shared resources strategy is consistent with the finding of "non-necessary" neurons in the frontal lobe, whose activity correlates with task performance, but which can be lesioned without noticeable effect on task performance (Tremblay et al., 2022). This result does not necessarily make sense from the point of view of optimal representation/computation or information theoretic efficiency. But it could fit into a routing framework. These neurons could be providing shared paths for relevant signals to traverse. In a distributed routing system with shared resources, no single router is strictly necessary, since signals can be actively rerouted. Removing one or several "non-necessary" nodes performing routing may not lead to a visible effect. In Tremblay et al. (2022) data, task performance-related activity peaks at different latencies in the "non-necessary" areas compared to "necessary" areas. This is consistent with a picture where different parts of the network are capable of flexibly transmitting the same messages over different paths.

Of course, much caution is due here. Without knowing network structure, results like Tremblay et al. (2022) can't on their own provide direct evidence for shared resources. Sharing may in fact be more important in the resource-limited and metabolically costly cortical white matter networks (Mollon and Danilova, 2019; Mollon et al., 2022) than in local cortical circuits. Despite the difficulty of recordings from intracortical white matter links at present (e.g., Li et al., 2016), a recognition of the importance of these signals as potentially evidence for shared resources could spur innovation in recording methods. In any case, evidence from human brain imaging of neural re-use (Anderson, 2010) and from neuroanatomy indicating computational flexibility (Pessoa et al., 2019) seems consistent with some level of shared resources in cortex.

Insight 4: Routing requires self-awareness

If signals have the ability to interfere with each other in a communication network that shares resources, each node would do well to exploit knowledge of its network environment to plan out a good route for messages it transmits. The system as a whole requires a kind of self-awareness—an on-going process for tracking network conditions and message deliveries. Internet routers monitor local network status to ensure they and their neighbors are aware of the existence of paths on shared links and current traffic load over those links. All devices wishing to join the network must support these core mechanisms of network monitoring. Mechanisms include *keep alives* which are regular heartbeat-like messages sent out by a router to all of its network neighbors to let them know the router is in service. There are also *echo requests*, which are small probe messages sent to a specific address, which must be reciprocated by the receiver, with all intermediaries reporting transit times for each leg of the journey. Perhaps most important of all are acknowledgments or *ACK*s, which are small return-receipt messages sent in retrograde fashion after a tranche of packets is successfully received. Note that these mechanisms are superfluous in computers: schematically, connectivity—e.g., two-way buses between processors and memory banks—is simple and highly reliable. Consider messages from processors to memory requesting stored data: the delivery of the data itself serves as confirmation that the request was received. Consequently, stand-alone computers do not generally require components to monitor each other or confirm signal receipt.

The brain, however, would seem to require systems for monitoring the operation of its communication network. Like the internet, such mechanisms would need to operate in distributed fashion over a network whose components are separated by comparatively long distances, suffer some degree of errors, and must trust each other.

Subnetworks in the brain could use spontaneous activity as a kind of keep alive-like message. In this scheme, spontaneous firing facilitates message passing along the same routes as those traveled by evoked signals. There is suggestive evidence of this. Mohajerani et al. (2013) used voltage sensitive dyes in the exposed cortex of mice, combined with prior connectomic maps, to show that both spontaneous and stimulus-evoked activity produced similar motifs of signal transmission. Mohajerani et al. (2013) call this pattern of spontaneous activity a "reverberation" of sensory signals, but perhaps it is better conceived as a preparation for transmitting such signals in the future. Spontaneous signals in this view serve as network status messages. Complementing these results is a microelectrode study in rat auditory cortex, Luczak et al. (2013), investigating what they called "packetization." Packets as defined in the study were repeating sequences of spike trains in different recorded neurons, much like the putative trajectories of Nádasdy et al. (1999), but rather unlike internet protocol packets. Luczak et al. (2013) found that spontaneous and stimulus evoked packets were similar in structure (see Luczak et al., 2015). This finding is consistent with the idea that neurons exchange content-bearing messages and network status messages on the same footing and over the same conduits. However, these findings are merely suggestive and do not serve as direct evidence of a keep-alive scheme. One intriguing avenue would be for experimentalists to test whether individual neurons or groups of neurons reliably pass signals on polysynaptic paths in ways that can be predicted based on prior patterns of spontaneous activity from the target of the path.

If patterns of transmission treat different types of messages (i.e., stimulus-evoked "content" and spontaneous network status signals) in similar ways, cortico-thalamic connection architecture would seem to naturally possess properties appropriate for providing delivery verification. Core brain networks would also appear to have a need for such a function. See Box 2 for a discussion of possible neuronal substrates in the thalamo-cortical networks that could support ACKs.

These hypothetical mechanisms for network status monitoring and delivery verification do not exactly mirror those used on the internet. Nevertheless, the metaphorical framework of the internet spurs us to conceptualize and investigate the brain in new and potentially transformative ways, which could help explain other puzzling problems at the core of brain organization. For example, the notion of a self-aware system of distributed, communicating elements offers a novel way to approach processes of allostasis in the brain (e.g., Sterling, 2012; Katsumi et al., 2022): the brain may need updates not only about the nature of planned or performed action but also knowledge of the network's readiness to carry out such actions.

Insight 5: Routing should be interoperable

Packaging all data into a standard size and structure, i.e., packets, not only allows sharing of resources, it also allows signals of different kinds—including messages with representational "content," as well as signals related to network status monitoring, and other kinds of messages—to travel together on the same network, all directed by the same routing rules. The potential for any imaginable data to be put into a packet was a basic part of ARPANET design, even though only two functions, remote login and file transfer, were possible on the original network, and indeed for decades afterward. Today this vision has been realized.

In the brain, we know there is a fundamental interoperability among cortical territories: for example, in sighted subjects, primary visual areas begin processing tactile stimuli within hours or days during blindfolding (Pascual-Leone and Hamilton, 2001), and this activity supports enhanced tactile sensitivity. This is not enough time to build extensive new connectivity—nor, presumably, to change the system's basic routing strategy. The influences of the messages of touch and their routing in vision processing systems were there all along and appear interoperable with vision-related signals in this part of the brain's communication network. Indeed, practically any real-world cognitive task requires integrating memories or knowledge from different domains (see e.g., Zeki, 2020). The requirement of interoperability applies not only between systems that deal in messages of different functional kinds but between systems of distinct phylogenetic ages, origins, and structure, such as cortical regions with six cell layers (isocortex/neocortex) and those with three or four layers (e.g., paleocortex).

Interoperability can be achieved in part by obviating the need to inspect or decode messages at most nodes. A router doesn't need to know what a packet contains. This is part of the cleverness of the internet: content is dealt with by senders and receivers, not by processing intermediaries. Could the same be true for neurons of the cortex? Consider that a "visual" neuron in V1 encoding an edge doesn't "know" about edges. Instead, it is responding based on inputs that traverse a particular network of connections before arriving at that neuron. However, its pattern of activity is often seen, under the computing/representational metaphor, as evidence that visual neurons do in fact "know" something about edges or faces or motion, because their spikes can be predicted fairly well from, e.g., deep learning models of visual representation (Yamins et al., 2014). We can draw inspiration from the design of internet routing to help us move beyond this kind of thinking. To complement the computer metaphor-based framework, I argue that we should start to consider things from the message's point of view: where a message originates, how it propagates and is transformed, how routing mechanisms deal with it and ensure it takes an efficient, reliable path, and where and when it is "delivered".

Insight 6: Routing should be scalable

The principles that govern internet routing are fully scalable: new links and nodes can be added gracefully, with modest cost to network operation. Network communication systems with topology and routing protocol that differ from those of the internet have less graceful scaling. Circuit-switched systems running on star-shaped networks, for example, risk overload without carefully planned growth: intuitively, if your neighbor adds a landline on a traditional telephone network, it will not affect communication over your landline. However, if too many new lines are added, switching stations risk running out of lines, preventing anyone not already using a line from starting a call. In contrast, the internet was specifically designed to scale up with modest cost and without central planning. Thanks to this design insight, most facets of internet routing strategy have required little fundamental modification even with rapid increases in nodes, links, and traffic.

The brain also undergoes upscaling in both ontogeny and phylogeny (though brains additionally experience network downscaling in the form of pruning and cell death). We should therefore expect a routing system in the brain that allows graceful scaling of message-passing; the system should by its nature avoid sharp discontinuities or precipitous changes, much as the internet does. It also must deal with the costs to network communication as it scales up.

Comparative investigations across mammals of different brain sizes could provide evidence of the costs of scaling up brain communication networks and could indicate a likely underlying strategy, just as can be done with black-box routing systems. As they scale up, brain divisions show consistent relationships between regional volume and overall volume (Finlay and Darlington, 1995), which are mediated by shifts in neurodevelopmental timelines. Neuron numbers, neuron size, neuron density, synapse numbers, and network topology scale together in more complex ways. However, the net effect of these relationships may have a global signature that reflects the brain's fundamental routing scheme. In particular, one could examine costs related to transmission of intrabrain signals. If these costs scale up monotonically but gradually in brains of increasing size, this would suggest an internet-like system that shares resources. In contrast, a routing system without shared resources—analogous to circuit-switched telephone networks—would show no increase in cost as the network grows since links are exclusive. However, such a system would be at risk of overload and could not scale organically. Cost in terms of metabolism may be difficult to define and measure but may be reflected in proxy measures. For example, the maximum speed of transmission of messages over multi-hop paths (normalized for distance traveled) could be such a proxy. All else being equal, maximum speed under an internet-like routing scheme would be hypothesized to slowly decrease (i.e., cost slowly increases) in bigger brains. If we see no slow down with brain size, this would be more consistent with a circuit-switched system. Other proxies such as sparseness may offer purchase on this question (see Graham, 2021). A detailed set of predictions about interrelationships among brain scaling, routing strategy and cost is outside the scope of the present paper but is under development.

Insight 7: Routing should be efficient

The internet would not have grown so gracefully if its basic operation had been too energetically expensive. It continues to succeed today despite massive network growth in part because message transport over optical fibers is very efficient (IEA, 2022). But beyond message transport, routing strategies implemented at nodes can also grant efficiency, sometimes in surprising ways.

Consider the back-off algorithm, a core tool found throughout internet-like networks. These algorithms deal with an ever-present problem: what to do when two messages "collide" i.e., attempt to occupy the same frame or clock tick at the input of, for example, an Ethernet router. When this happens, both messages are destroyed. For each destroyed message the router then essentially flips a coin—heads, a copy of the message cached at its sender is allowed pass, tails, it has to wait for the next tick. If a message collides on further attempts and draws tails, it has to wait up to 2 ticks, then up to 4 ticks, then up to 8 ticks. This algorithm is termed binary exponential back-off. It results in an exponential distribution of delays. The basic design principle of imposing randomized wait times for colliding messages has been in place since the earliest days of internet-based communication systems, starting with the ALOHA packet-switched radio network in Hawaii (Kleinrock, 1976). Routinely injecting timing noise into message passing systems remains a cornerstone of routing efficiency across the internet. Notice that this is an example of an engineering insight in communication that differs greatly from insights exploited in the computer metaphor: adding timing noise at all nodes would not have an obvious benefit for representational systems (though deep learning systems do employ "drop-out" for somewhat related purposes) but is demonstrably successful when communication is the goal.

The possibility of something like exponential back-off in the brain is worth consideration. The ubiquity of Poisson-like spike generation in the mammalian cerebral cortex (see, e.g., Averbeck, 2009) produces exponentially-distributed interspike intervals (ISIs). If we see ISIs as delays, this behavior is consistent with the brain performing exponential back-off as part of its network communication strategy. If the brain uses similar routing strategies as those described in this paper, exponentially-distributed ISIs could serve to minimize the effects of collisions. Though many processes generate exponential distributions, this distribution is a hallmark of internet dynamics, so it is curious that a similar distribution is found also throughout mammal cortex. Exponentially-distributed ISIs are observed also for spontaneous firing in cortex (Mazzoni et al., 2007) suggesting that they are not due only to dynamics of stimulus experience but also due to intrinsic factors. However, back-off algorithms, like ACKs, require node buffers to allow resending, which remain hypothetical in the brain (see Box 2).

We can take a wider view of efficiency. Progress in understanding representational aspects of brain function has been aided by efficiency arguments (e.g., Doi and Lewicki, 2014), so a similar approach may be profitable in terms of communication. It has long been clear that transmitting signals down axons is very costly, leading to wiring minimization models (see e.g., Cherniak et al., 2002; Chklovskii and Koulakov, 2004). A recent estimate suggests neuronal communication is in fact far more costly than neuronal computation (Balasubramanian, 2021; Levy and Calvert, 2021). With transmission already expensive, routing strategies in the brain—whose costs were not considered in the estimate of Levy and Calvert (2021)—must be shaped to a significant degree by efficiency concerns. Sparse activity in time and space also contributes to efficiency both on the internet and in the brain; (see Graham, 2021; Graham and Rockmore, 2011) for further discussion of the role of sparseness.

However, gauging the large-scale efficiency of routing in the brain will be a major challenge because we lack a mathematical formalism for describing information theoretic limits on network communication. Shannon's information theory, which is widely invoked in studies of efficiency in neuronal representations (e.g., Wainwright, 1999) applies only to the case of "two-port" communication (Cover and Thomas, 1991), i.e., point encoding and decoding with channel noise. New approaches to the study of efficiency in network communication broadly construed may be needed [see the "network information theory" of El Gamal and Kim (2011); see also Pastor-Satorras and Vespignani (2004), Sun et al. (2015); and Amico et al. (2021)]. The problem in the network case is that resources are shared. One needs to balance cost and reliability when ongoing signal generation can influence individual routing actions in very complex ways. The efficient solutions the brain employs—if they can be determined, and if they are in some sense optimal—may in fact help us understand more fundamental principles of network information theory. However, the internet's demonstrable efficiency suggests that basic principles of efficient network communication may already be instantiated in its array of solutions.

Insight 8: Asynchronous routing can simulate synchrony

Distributed strategies have advantages but also impose constraints: the internet, for example, is fundamentally an asynchronous communication system: senders and receivers generally cannot establish complete, on-going circuits; full "communion" is unachievable. However, because path lengths are short and because delays on the network are miniscule by human standards, a simulation of synchrony is possible. Human senders and receivers readily perceive its real-time functions (e.g., video chat) as synchronous and simultaneous.[5]

In the brain, anatomical and network distances are long enough and propagation of neuronal excitations slow enough that the system as a whole functions asynchronously, even if our conscious experience makes it feel as if there is a fully synchronous, delay-free "now" (see, e.g., Zeki and Bartels, 1998; Hogendoorn, 2021). However, functions like object perception may achieve short intervals of synchrony on subnetworks, allowing faster and more coordinated action among dispersed brain elements (Gray et al., 1989; Fries, 2005; Vezoli et al., 2021; Uran et al., 2022).[6] Elaborate systems of precisely-timed delays are also a basic feature of cortical signaling, which helps coordinate activity of asynchronous elements (Innocenti et al., 2016).

Seemingly, routing in the brain, as on the internet, is fundamentally asynchronous, but is capable of simulating synchrony over short time scales among subgroups of nodes. The internet's solutions, such as content delivery networks used by Netflix and others, which store multiple copies of a given resource at different points on the network, are worth considering in reference to brain networks.

Insight 9: Routing should be unified, but can be modified locally

Despite being composed of billions of dispersed elements, the internet is a unified entity. However, unity is not imposed from a central controller. Instead, a common set of routing rules

[5] However, even the short time delays of the internet can become noticeable. In a Zoom meeting, try this experiment: One person starts to clap at a slow tempo. Then others try to match the beat. Participants often become 180 degrees out of phase with the reference clap.

[6] The idea of synchrony in digital computation and communication systems is not entirely equivalent to synchrony in dynamical systems of the kind described by Vezoli et al. (2021) and others. In dynamical systems, synchrony usually implies periodicity and occurs when oscillators coordinate the timing of their actions. In digital computing, synchrony means that one system can interfere with the concurrent operation of another system, but there is not necessarily periodicity.

is implemented locally, and the rules can be locally modified to some extent as well. For example, subnetworks can prioritize some packets over others, and novel devices and protocol can be added so long as basic protocol is followed. A few network services on the internet are organized centrally (e.g., the domain name server) but this is largely a convenience for human end-users. Basic operations of message transmission require no central entity (nor human intervention: connectable devices can now be found and added to the network automatically; Mišić and Mišić, 2014). Even key services like time keeping, which is performed centrally in a computer, are decentralized on the internet using network time protocol. The internet does include modular elements (e.g., autonomous systems), which can exert specialized, central control over a domain (e.g., firewalls). But each module must ultimately be compatible with the wider network by way of common routers. And in the operation of an autonomous system (AS), most of the same "tricks" found in parts of the network outside the AS are employed internally as well.

Protocol in the brain may likewise be a global phenomenon where a relatively small set of rules apply equally throughout, but can be modified. It is reasonable to first consider if there is a single protocol for the whole mammal or vertebrate brain. A patchwork of multiple, non-overlapping networks, each with its own protocol, seems more characteristic of nerve nets (Dupre and Yuste, 2017) than of highly structured and interconnected brains like those of mammals, birds, and especially humans (Hofman, 1988). But we should not rule out the possibility of overlapping neural systems running on different protocol that achieve widespread influence on the network but with limited functionality, such as the fast emergency alert system centered on brainstem nuclei.

Unified protocol in the brain could tolerate considerable local variation and tuning in different regions. Different species may also show specialization in routing. Local modification of global routing rules may influence brain organization within a species. The cortex is typified by "unity and diversity" of structure, shown in, for example, its laminar and columnar architecture (Schüz and Miller, 2002). In different brain regions, variations in a set of conserved genes could shape overall routing strategies. Through the effect of interactions of these genes, small tweaks in units controlling how routing "protocol" is implemented during neurodevelopment could generate significant changes in brain dynamics and function, just as small tweaks to cell growth can affect brain size (e.g., "late equals large" Finlay et al., 2010). With local alteration in routing could come a diversity of function. Through such mechanisms, distributed specializations of protocol in different brain systems could be engineered without sacrificing global integration. The same basic process may help shape phylogenetic variation.

These insights allow us to imagine a rich problem space that we can consider in relation to the study of the brain's strategy for network communication. Internet engineering provides a collection of effective strategies that may be similar to those the brain uses. However, a full description of the neuronal toolkit that could implement the above functions is needed. Testable hypotheses will need to be developed, and these will require more precise models of possible neuronal substrates than I have offered here.

Discussion

A good metaphor in scientific theory indicates the span and orientation of a problem to be solved. Like a microscope or an electrode, metaphor is a tool, one used in service of theory, rather than experiment. Metaphor is not sufficient for theory, but can be its precursor. Metaphor can help us get to a place where we can specify quantities of interest and understand why measuring those quantities and not others will be meaningful. Technological metaphor, because it refers to engineered systems with goals, is of special potential use since biological organisms and their brains are shaped by evolutionary "engineering." We can ask, "what would be a good way to design a neuronal system that must operate under certain conditions, such as those that permit flexible exchange of signals across a densely connected network?" The routing strategies of the internet, a technological system that was specifically designed to solve such problems, are worth consideration in relation to this question.

Yet if we grant that the brain must perform flexible routing, the endeavor to understand its strategies in light of the internet still faces major difficulties. One is addressing, a key feature of any routing scheme, which shapes all other features. Selective communication presupposes the existence of an addressing system, though explicit address "headers" may not be required in the brain. Schemes that invoke synchronous oscillations (e.g., Fries, 2005; Nádasdy, 2010) seem to obviate the need for "headers" that travel with a message, but such schemes have not yet dealt with how targets are selectively chosen, nor how congestion could be managed. Indeed, these problems have not been recognized. If headers are needed, spike timing could conceivably carry this information: most paths are likely to be only a few hops in length, so header information could be small. However, detailed models of this kind have not been elucidated let alone tested.

Metaphors can be misleading, especially if taken too literally. This is true not least for the computer metaphor. A physical computer, unlike the brain, has a clock that strictly synchronizes all operations, while a Turing machine requires infinite "tape" on which to order symbols. The brain is obviously not a literal internet either.

Nor can we say that the internet is the only good solution to the problem of dynamic network communication. There exist an unknown number of possible schemes. One can imagine communication systems that include multiple senders and/or receivers where a "multi-message" of distributed chunks travels on parallel paths; along similar lines, it may be better to think in terms of "sources" and "sinks" of signal flow (Mohajerani et al., 2013) rather than single copies of messages with a single path. Part of the problem here is that we lack a grounding of network communication theory in terms of basic mathematics. This is a contrast with the view of brains as computers and representation machines, where we understand the fundamental limitations and possibilities thanks to the well-understood underlying theories of functions and computability.

But the internet's demonstrable success—through pandemic, war, and malicious attack—suggests it embodies basic insights regarding the organization and integration of flexible message flow on large, complex, growing networks. Ultimately, a turn toward the internet metaphor accords with the longstanding desire to understand the integration of computational functions in the brain, and how distributed signals are related and bound to one another (e.g., Popper and Eccles, 1977). The internet metaphor offers more precise language and deeper analogies compared to earlier analogies of brain integration, such as "workspaces" (Dehaene and Changeux, 2005; Baldauf and Deubel, 2010), "bulletin boards," (Baars, 1997; Goyal et al., 2021), "puzzle pieces" (Chater, 2018; John et al., 2022), or reactions involving "catalysts" and "bonding" (Varela et al., 1991). Brain science stands to profit from considering the internet's strategies and solutions and from asking how the brain might solve similar problems. An understanding of routing in the brain has the potential to illuminate many aspects of brains, not least the decipherment of neural codes, but also evolutionary and developmental patterns, functional differentiation, neurological conditions affecting large-scale brain intra-communication (e.g., multiple sclerosis and epilepsy), as well as intelligence and consciousness.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

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

Acknowledgments

I would like to thanks David Field, Barbara Finlay, Don Spector, Yan Hao, Marc-Thorsten Hütt, Kevin Tang, and Niko Grupen for helpful discussions.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Amico, E., Abbas, K., Duong-Tran, D. A., Tipnis, U., Rajapandian, M., Chumin, E., et al. (2021). Toward an information theoretical description of communication in brain networks. *Netw. Neurosci.* 5, 646–665. doi: 10.1162/netn_a_00185

Anderson, J. C., and Martin, K. A. C. (2016). "Interareal connections of the macaque cortex: how neocortex talks to itself," in *Axons and brain architecture* (Academic Press), 117–134. doi: 10.1016/B978-0-12-801393-9.00006-2

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behav. Brain Sci.* 33, 245–266. doi: 10.1017/S0140525X10000853

Anderson, M. L., and Champion, H. (2022). Some dilemmas for an account of neural representation: A reply to Poldrack. *Synthese.* 200, 1–13. doi: 10.1007/s11229-022-03505-4

Avena-Koenigsberger, A., Miši,ć, B., Hawkins, R. X., Griffa, A., Hagmann, P., Goñi, J., et al. (2017). Path ensembles and a tradeoff between communication efficiency and resilience in the human connectome. *Brain Struct. Funct.* 222, 603–618. doi: 10.1007/s00429-016-1238-5

Averbeck, B. B. (2009). Poisson or not poisson: differences in spike train statistics between parietal cortical areas. *Neuron* 62, 310–311. doi: 10.1016/j.neuron.2009.04.021

Baars, B. J. (1997). *In the Theater of Consciousness: The Workspace of the Mind.* USA: Oxford University Press. doi: 10.1093/acprof:oso/9780195102659.001.1

Baker, B., Lansdell, A., and Kording, K. (2022). Three aspects of representation in neuroscience. *Trends Cognit. Sci.* 26, 942–958. doi: 10.1016/j.tics.2022.08.014

Balasubramanian, V. (2021). Brain power. *Proc. Nat. Acad. Sci.* 118, e2107022118. doi: 10.1073/pnas.2107022118

Baldauf, D., and Deubel, H. (2010). Attentional landscapes in reaching and grasping. *Vision research* 50, 999–1013. doi: 10.1016/j.visres.2010.02.008

Barabási, A. L., and Albert, R. (1999). Emergence of scaling in random networks. *science* 286, 509–512. doi: 10.1126/science.286.5439.509

Baran, P. (1964). On distributed communications networks. *IEEE transactions on Communications Systems* 12, 1–9. doi: 10.1109/TCOM.1964.1088883

Bargmann, C. I., and Marder, E. (2013). From the connectome to brain function. *Nature methods* 10, 483–490. doi: 10.1038/nmeth.2451

Bartha, P. (2022). "Analogy and Analogical Reasoning", *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/sum2022/entries/reasoning-analogy/>.

Boehm, S. P., and Baran, P. (1964). *On distributed* communications: II. Digital simulation of hot-potato routing in a broadband distributed communications network. Memorandum of the RAND corporation prepared for United States Air Force.

Boroujeni, K. B., and Womelsdorf, T. (2022). Routing States Transition During Oscillatory Bursts and Attention States *bioRxiv* 2022.10.29.514374.

Brette, R. (2019). Is coding a relevant metaphor for the brain?. *Behavioral and Brain Sciences* 42. doi: 10.1017/S0140525X19000049

Brette, R. (2022). Brains as computers: metaphor, analogy, theory or fact? *Frontiers in Ecology and Evolution* 10, 878729. doi: 10.3389/fevo.2022.878729

Briggs, F., and Usrey, W. M. (2007). A Fast, Reciprocal Pathway Between the Lateral Geniculate Nucleus and Visual Cortex in the Macaque Monkey. *Journal of Neuroscience* 27, 20: 5431–5436. doi: 10.1523/JNEUROSCI.1035-07.2007

Bruineberg, J., Dolega, K., Dewhurst, J., and Baltieri, M. (2021). The Emperor's New Markov Blankets. *Behavioral and Brain Sciences,* 1-63. doi: 10.1017/S0140525X21002351

Buzsáki, G. (2004). Large-scale recording of neuronal ensembles. *Nature neuroscience* 7, 446–451. doi: 10.1038/nn1233

Buzsáki, G. (2019). *The brain from inside out*. Oxford University Press. doi: 10.1093/oso/9780190905385.001.0001

Caldarelli, G., Marchetti, R., and Pietronero, L. (2000). The fractal properties of Internet. *EPL (Europhysics Letters)* 52, 386. doi: 10.1209/epl/i2000-00450-8

Chater, N. (2018). *The mind is* flat: The remarkable shallowness of the improvising brain. Yale University Press. doi: 10.12987/9780300240610

Cherniak, C., Mokhtarzada, Z., and Nodelman, U. (2002). Optimal-wiring models of neuroanatomy. *Computational neuroanatomy* 71–82. doi: 10.1385/1-59259-275-9:71

Chklovskii, D. B., and Koulakov, A. A. (2004). Maps in the brain: what can we learn from them?. *Annual review of neuroscience* 27, 369–392. doi: 10.1146/annurev.neuro.27.070203.144226

Cobb, M. (2020). *The idea of the brain: The past and future of neuroscience.* Hachette UK.

Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., and Braver, T. S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature neuroscience* 16, 1348–1355. doi: 10.1038/nn.3470

Colizza, V., Flammini, A., Serrano, M. A., and Vespignani, A. (2006). Detecting rich-club ordering in complex networks. *Nature physics* 2, 110–115. doi: 10.1038/nphys209

Cover, T., and Thomas, J. (1991). *Elements of Information Theory*. New York, NY: Wiley. doi: 10.1002/0471200611

Crabtree, J. W. (2018). Functional Diversity of Thalamic Reticular Subnetworks. *Front Syst Neurosci.* 12:41. doi: 10.3389/fnsys.2018.00041

Danilova, M. V., and Mollon, J. D. (2003). Comparison at a distance. *Perception* 32, 395–414. doi: 10.1068/p3393

Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The "neuronal recycling" hypothesis. *From monkey brain to human brain* 133–157. doi: 10.7551/mitpress/3136.003.0012

Dehaene, S., and Changeux, J. P. (2005). Ongoing spontaneous activity controls access to consciousness: a neuronal model for inattentional blindness. *PLoS Biol.* 3, e141. doi: 10.1371/journal.pbio.0030141

Doi, E., and Lewicki, M. S. (2014). A simple model of optimal population coding for sensory systems. *PLoS computational biology* 10, e1003761. doi: 10.1371/journal.pcbi.1003761

Doyle, J. C., Alderson, D. L., Li, L., Low, S., Roughan, M., Shalunov, S., et al. (2005). The "robust yet fragile" nature of the Internet. *Proc. Nat. Acad. Sci.* 102, 14497–14502. doi: 10.1073/pnas.0501426102

Dupre, C., and Yuste, R. (2017). Non-overlapping neural networks in Hydra vulgaris. *Curr. Biol.* 27, 1085–1097. doi: 10.1016/j.cub.2017.02.049

El Gamal, A., and Kim, Y. H. (2011). *Network Information Theory*. USA: Cambridge University Press. doi: 10.1017/CBO9781139030687

Epsztein, J., Lee, A. K., Chorev, E., and Brecht, M. (2010). Impact of spikelets on hippocampal CA1 pyramidal cell activity during spatial exploration. *Science* 327, 474–477. doi: 10.1126/science.1182773

Fauth, M., and Tetzlaff, C. (2016). Opposing effects of neuronal activity on structural plasticity. *Front. Neuroanat.* 10, 75. doi: 10.3389/fnana.2016.00075

Fields, C., Glazebrook, J. F., and Levin, M. (2022). Neurons as hierarchies of quantum reference frames. *Biosystems.* 129, 104714. doi: 10.1016/j.biosystems.2022.104714

Finlay, B. L., Clancy, B., and Darlington, R. B. (2010). Late still equals large. *Brain, Behav. Evolut.* 75, 4. doi: 10.1159/000295350

Finlay, B. L., and Darlington, R. B. (1995). Linked regularities in the development and evolution of mammalian brains. *Science* 268, 1578–1584. doi: 10.1126/science.7777856

Fornito, A., Zalesky, A., and Breakspear, M. (2015). The connectomics of brain disorders. *Nat. Rev. Neurosci.* 16, 159–172. doi: 10.1038/nrn3901

Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends Cognit. Sci.* 9, 11. doi: 10.1016/j.tics.2005.08.011

Fukushima, M., and Leibnitz, K. (2022). Packetization improves communication efficiency in brain networks with rapid and cost-effective propagation strategies. *bioRxiv*. doi: 10.1101/2022.06.30.498099

Funahashi, S. (2015). Functions of delay-period activity in the prefrontal cortex and mnemonic scotomas revisited. *Front. Syst. Neurosci.* 9, 2. doi: 10.3389/fnsys.2015.00002

Gămănut, R., Kennedy, H., Toroczkai, Z., Ercsey-Ravasz, M., Van Essen, D. C., Knoblauch, K., et al. (2018). The mouse cortical connectome, characterized by an ultra-dense cortical graph, maintains specificity by distinct connectivity profiles. *Neuron* 97, 698–715. doi: 10.1016/j.neuron.2017.12.037

Gerraty, R. T., Davidow, J. Y., Foerde, K., Galvan, A., Bassett, D. S., and Shohamy, D. (2018). Dynamic flexibility in striatal-cortical circuits supports reinforcement learning. *J. Neurosci.* 38, 2442–2453. doi: 10.1523/JNEUROSCI.2084-17.2018

Gidon, A., Aru, J., and Larkum, M. (2022). Does brain activity cause consciousness? A thought experiment. *PLoS Biol.* 20, e3001651. doi: 10.1371/journal.pbio.3001651

Gidon, A., Zolnik, T. A., Fidzinski, P., Bolduan, F., Papoutsi, A., Poirazi, P., et al. (2020). Dendritic action potentials and computation in human layer 2/3 cortical neurons. *Science* 367, 83–87. doi: 10.1126/science.aax6239

Gisiger, T., and Boukadoum, M. (2011). Mechanisms gating the flow of information in the cortex: what they might look like and what their uses may be. *Front. Comput. Neurosci.* 5, 1. doi: 10.3389/fncom.2011.00001

Gold, J. I., and Shadlen, M. N. (2007). The neural basis of decision making. *Ann. Rev. Neurosci.* 30, 535–574. doi: 10.1146/annurev.neuro.29.051605.113038

Goldman-Rakic, P. S. (1996). Regional and cellular fractionation of working memory. *Proc. Nat. Acad. Sci.* 93, 13473–13480. doi: 10.1073/pnas.93.24.13473

Gollisch, T., and Meister, M. (2010). Eye smarter than scientists believed: neural computations in circuits of the retina. *Neuron* 65, 150–164. doi: 10.1016/j.neuron.2009.12.009

Goyal, A., Didolkar, A., Lamb, A., Badola, K., Ke, N. R., Rahaman, N., et al. (2021). Coordination among neural modules through a shared global workspace. arXiv preprint arXiv:2103.01197.

Graham, D. (2021). *An Internet in Your Head: A New Paradigm for How the Brain Works.* New York, NY: Columbia University Press. doi: 10.7312/grah19604

Graham, D., Avena-Koenigsberger, A., and Mišić, B. (2020). Network communication in the brain. *Netw. Neurosci.* 4, 976–979. doi: 10.1162/netn_e_00167

Graham, D., and Rockmore, D. (2011). The packet switching brain. *J. Cognit. Neurosci.* 23, 267–276. doi: 10.1162/jocn.2010.21477

Graham, D. J. (2014). Routing in the brain. *Front. Comput. Neurosci.* 8, 44. doi: 10.3389/fncom.2014.00044

Gray, C. M., König, P., Engel, A. K., and Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* 338, 334–337. doi: 10.1038/338334a0

Griffiths, T. L., Steyvers, M., and Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. *Psychol. Sci.* 18, 1069–1076. doi: 10.1111/j.1467-9280.2007.02027.x

Grosmark, A. D., and Buzsáki, G. (2016). Diversity in neural firing dynamics supports both rigid and learned hippocampal sequences. *Science* 351, 1440–1443. doi: 10.1126/science.aad1935

Günseli, E., and Aly, M. (2020). Preparation for upcoming attentional states in the hippocampus and medial prefrontal cortex. *Elife* 9, e53191. doi: 10.7554/eLife.53191.sa2

Guo, Z. V., Inagaki, H. K., Daie, K., Druckmann, S., Gerfen, C. R., and Svoboda, K. (2017). Maintenance of persistent activity in a frontal thalamocortical loop. *Nature* 545, 181–186. doi: 10.1038/nature22324

Hao, Y., and Graham, D. (2020). Creative destruction: Sparse activity emerges on the mammal connectome under a simulated communication strategy with collisions and redundancy. *Netw. Neurosci.* 4, 1055–1071. doi: 10.1162/netn_a_00165

Hilgetag, C. C., Medalla, M., Beul, S. F., and Barbas, H. (2016). The primate connectome in context: principles of connections of the cortical visual system. *NeuroImage* 134, 685–702. doi: 10.1016/j.neuroimage.2016.04.017

Hipólito, I. (2022). Cognition without neural representation: dynamics of a complex system. *Front. Psychology* 5472. doi: 10.3389/fpsyg.2021.643276

Hodassman, S., Vardi, R., Tugendhaft, Y., Goldental, A., and Kanter, I. (2022). Efficient dendritic learning as an alternative to synaptic plasticity hypothesis. *Scientific Rep.* 12, 1–12. doi: 10.1038/s41598-022-10466-8

Hofman, M. A. (1988). Size and shape of the cerebral cortex in mammals. *Brain, Behav. Evolut.* 32, 17–26. doi: 10.1159/000116529

Hogendoorn, H. (2021). Perception in real-time: predicting the present, reconstructing the past. *Trends Cogn. Sci.* 26, 128–141. doi: 10.1016/j.tics.2021.11.003

IEA (2022). *Data Centres and Data Transmission Networks*, IEA, Paris. License: CC *BY* 4.0. Available online at: https://www.iea.org/reports/data-centres-and-data-transmission-networks (accessed December 21, 2022).

Innocenti, G. M., Carlén, M., and Dyrby, T. B. (2016). "The Diameters of Cortical Axons and Their Relevance to Neural Computing in Axons and Brain Architecture," in *Axons and Brain Architecture,* ed. R. Kathleen (New York, NY: Academic Press) 317–355. doi: 10.1016/B978-0-12-801393-9.00015-3

Javadzadeh, M., and Hofer, S. B. (2021). Dynamic causal communication channels between neocortical areas. *Neuron* 110, 1–14. doi: 10.1101/2021.06.28.449892

John, Y. (2022). 'Representing' means exactly what you think it means. Available online at: https://yohanjohn.com/neurologism/representing-means-exactly-what-you-think-it-means/ (accessed October 7, 2022).

John, Y. J., Sawyer, K. S., Srinivasan, K., Müller, E. J., Munn, B. R., and Shine, J. M. (2022). It's about time: Linking dynamical systems with human neuroimaging to understand the brain. *Netw. Neurosci.* 6, 960–979. doi: 10.1162/netn_a_00230

Katsumi, Y., Theriault, J. E., Quigley, K. S., and Barrett, L. F. (2022). Allostasis as a core feature of hierarchical gradients in the human brain. *Netw. Neurosci.* 6, 1010–1031. doi: 10.1162/netn_a_00240

Kleinrock, L. (1976). *Queueing Systems, Vol II: Computer Applications*. New York, NY: Wiley.

Knoblauch, K., Ercsey-Ravasz, M., Kennedy, H., and Toroczkai, Z. (2016). "The brain in space," in *Micro-, meso-and macro-connectomics of the Brain,* 45–74. doi: 10.1007/978-3-319-27777-6_5

Kreiter, A. K. (2020). Synchrony, flexible network configuration, and linking neural events to behavior. *Curr. Opin. Physiol.* 16, 98–108. doi: 10.1016/j.cophys.2020.08.008

Latora, V., and Marchiori, M. (2001). Efficient behavior of small-world networks. *Phys. Rev. Lett.* 87, 198701. doi: 10.1103/PhysRevLett.87.198701

Levy, W. B., and Calvert, V. G. (2021). Communication consumes 35 times more energy than computation in the human cortex, but both costs are needed to predict synapse number. *Proc. Nat. Acad. Sci.* 118, 173118. doi: 10.1073/pnas.2008173118

Li, L., Alderson, D., Willinger, W., and Doyle, J. (2004). A first-principles approach to understanding the internet's router-level topology. *ACM SIGCOMM Comput. Commun. Rev.* 34, 3–14. doi: 10.1145/1030194.1015470

Li, L., Velumian, A. A., Samoilova, M., and Fehlings, M. G. (2016). A novel approach for studying the physiology and pathophysiology of myelinated and non-myelinated axons in the CNS white matter. *PLoS ONE* 11, e0165637. doi: 10.1371/journal.pone.0165637

Lindsay, G. (2021). *Models of the Mind: How Physics, Engineering and Mathematics Have Shaped Our Understanding of the Brain.* London: Bloomsbury Publishing. doi: 10.5040/9781472966445

Luczak, A., Bartho, P., and Harris, K. D. (2013). Gating of sensory input by spontaneous cortical activity. *J. Neurosci.* 33, 1684–1695. doi: 10.1523/JNEUROSCI.2928-12.2013

Luczak, A., McNaughton, B. L., and Harris, K. D. (2015). Packet-based communication in the cortex. *Nat. Rev. Neurosci.* 16, 745–755. doi: 10.1038/nrn4026

Matsuyama, K., and Tanaka, M. (2021). Temporal prediction signals for periodic sensory events in the primate central thalamus. *J. Neurosci.* 41, 1917–1927. doi: 10.1523/JNEUROSCI.2151-20.2021

Mayner, W. G., Marshall, W., Billeh, Y. N., Gandhi, S. R., Caldejon, S., Cho, A., et al. (2022). Measuring stimulus-evoked neurophysiological differentiation in distinct populations of neurons in mouse visual cortex. *Eneuro.* 9, 280. doi: 10.1523/ENEURO.0280-21.2021

Mazzoni, A., Broccard, F. D., Garcia-Perez, E., Bonifazi, P., Ruaro, M. E., and Torre, V. (2007). On the dynamics of the spontaneous activity in neuronal networks. *PLoS ONE* 2, e439. doi: 10.1371/journal.pone.0000439

McCulloch, W. S., and Pfeiffer, J. (1949). Of digital computers called brains. *Scientific Monthly* 69, 368–376.

Mehler, D. M. A., and Kording, K. P. (2018). The lure of misleading causal statements in functional connectivity research. arXiv preprint arXiv:1812.03363.

Meyers, M. (2004). *Network+ Certification Exam Guide*. New York, NY: McGraw Hill Professional.

Mishra, J., Fellous, J. M., and Sejnowski, T. J. (2006). Selective attention through phase relationship of excitatory and inhibitory input synchrony in a model cortical neuron. *Neural Netw.* 19, 1329–1346. doi: 10.1016/j.neunet.2006.08.005

Mišić, B., Betzel, R. F., Nematzadeh, A., Goi, J., Griffa, A., Hagmann, P., et al. (2015). Cooperative and competitive spreading dynamics on the human connectome. *Neuron* 86, 1518–1529. doi: 10.1016/j.neuron.2015.05.035

Mišić, B., Sporns, O., and McIntosh, A. R. (2014). Communication efficiency and congestion of signal traffic in large-scale brain networks. *PLoS Comput. Biol.* 10, e1003427. doi: 10.1371/journal.pcbi.1003427

Mišić, V. B., and Mišić, J. (2014). *Machine-to-Machine Communications: Architectures, Technology, Standards, and Applications.* Boca Raton, FL: CRC Press.

Mohajerani, M. H., Chan, A. W., Mohsenvand, M., LeDue, J., Liu, R., McVea, D. A., et al. (2013). Spontaneous cortical activity alternates between motifs defined by regional axonal projections. *Nat. Neurosci.* 16, 1426–1435. doi: 10.1038/nn.3499

Möller, C., Lücke, J., Zhu, J., Faustmann, P. M., and Von Der Malsburg, C. (2007). Glial cells for information routing? *Cogn. Syst. Res.* 8, 28–35. doi: 10.1016/j.cogsys.2006.07.001

Mollon, J., and Danilova, M. (2019). Cortical communication and the comparison of colors. *Curr. Opin. Behav. Sci.* 30, 203–209. doi: 10.1016/j.cobeha.2019.10.002

Mollon, J. D., Takahashi, C., and Danilova, M. V. (2022). What kind of network is the brain? *Trends Cogn. Sci.* 26, 312–324. doi: 10.1016/j.tics.2022.01.007

Nádasdy, Z. (2010). Binding by asynchrony: the neuronal phase code. *Front. Neurosci.* 4, 51. doi: 10.3389/fnins.2010.00051

Nádasdy, Z., Hirase, H., Czurkó, A., Csicsvari, J., and Buzsáki, G. (1999). Replay and time compression of recurring spike sequences in the hippocampus. *J. Neurosci.* 19, 9497–9507. doi: 10.1523/JNEUROSCI.19-21-09497.1999

Navlakha, S., Bar-Joseph, Z., and Barth, A. L. (2018). Network design and the brain. *Trends Cogn. Sci.* 22, 64–78. doi: 10.1016/j.tics.2017.09.012

Nelson, M. E., and Bower, J. M. (1990). Brain maps and parallel computers. *Trends Neurosci.* 13, 403–408. doi: 10.1016/0166-2236(90)90119-U

Oka, M., Abe, H., and Ikegami, T. (2015). Dynamic homeostasis in packet switching networks. *Adapt. Behav.* 23, 50–63. doi: 10.1177/1059712314556369

Olshausen, B. A., Anderson, C. H., and Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* 13, 4700–4719. doi: 10.1523/JNEUROSCI.13-11-04700.1993

Olshausen, B. A., and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0

Oz, O., Matityahu, L., Mizrahi-Kliger, A., Kaplan, A., Berkowitz, N., Tiroshi, L., et al. (2021). Non-uniform distribution of dendritic nonlinearities differentially engages thalamostriatal and corticostriatal inputs onto cholinergic interneurons. *bioRxiv*. doi: 10.1101/2021.11.29.470423

Palmigiano, A., Geisel, T., Wolf, F., and Battaglia, D. (2017). Flexible information routing by transient synchrony. *Nat. Neurosci.* 20, 1014–1022. doi: 10.1038/nn.4569

Parsons, N., Ugon, J., Morgan, K., Shelyag, S., Hocking, A., Chan, S. Y., et al. (2022). Structural-functional connectivity bandwidth of the human brain. *NeuroImage* 263, 119659. doi: 10.1016/j.neuroimage.2022.119659

Pascual-Leone, A., and Hamilton, R. (2001). The metamodal organization of the brain. *Progr. Brain Res.* 134, 427–445. doi: 10.1016/S0079-6123(01)34028-1

Pastor-Satorras, R., and Vespignani, A. (2004). *Evolution and Structure of the Internet: A Statistical Physics Approach.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511610905

Pavlov, I. P. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex, trans.* G. V. Anrep Oxford: Oxford University Press.

Pessoa, L., Medina, L., Hof, P. R., and Desfilis, E. (2019). Neural architecture of the vertebrate brain: implications for the interaction between emotion and cognition. *Neurosci. Biobehav. Rev.* 107, 296–312. doi: 10.1016/j.neubiorev.2019.09.021

Poggio, T. (1984). *Routing thoughts*. Massachusetts Institute of Technology Artificial Intelligence Laboratory Working Paper 258.

Poldrack, R. A. (2021). The physics of representation. *Synthese* 199, 1307–1325. doi: 10.1007/s11229-020-02793-y

Popper, K. R., and Eccles, J. C. (1977). *The Self and its Brain*. Berlin: Springer International. doi: 10.1007/978-3-642-61891-8

Quammen, D. (2018). *The Tangled Tree: A Radical New History of Life.* New York, NY: Simon and Schuster.

Reichova, I., and Sherman, S. M. (2004). Somatosensory corticothalamic projections: Distinguishing drivers from modulators. *J. Neurophysiol.* 92, 2185–2197. doi: 10.1152/jn.00322.2004

Richards, B. A., and Lillicrap, T. P. (2022). The brain-computer metaphor debate is useless: A matter of semantics. *Front. Comput. Sci.* 4, 810358. doi: 10.3389/fcomp.2022.810358

Rosvall, M., Trusina, A., Minnhagen, P., and Sneppen, K. (2005). Networks and cities: An information perspective. *Phys. Rev. Lett.* 94, 028701. doi: 10.1103/PhysRevLett.94.028701

Safron, A., Klimaj, V., and Hipólito, I. (2022). On the importance of being flexible: dynamic brain networks and their potential functional significances. *Front. Syst. Neurosci.* 149, 688424. doi: 10.3389/fnsys.2021.688424

Sakalar, E., Klausberger, T., and Lasztóczi, B. (2022). Neurogliaform cells dynamically decouple neuronal synchrony between brain areas. *Science* 377, 324–328. doi: 10.1126/science.abo3355

Schüz, A., and Miller, R. (2002). *Cortical Areas: Unity and Diversity.* London: CRC press.

Scott, A. (1977). *Neurophysics*. New York, NY: John Wiley.

Seguin, C., Razi, A., and Zalesky, A. (2019). Inferring neural signalling directionality from undirected structural connectomes. *Nat. Commun.* 10, 1–13. doi: 10.1038/s41467-019-12201-w

Seguin, C., Van Den Heuvel, M. P., and Zalesky, A. (2018). Navigation of brain networks. *Proc. Nat. Acad. Sci.* 115, 6297–6302. doi: 10.1073/pnas.1801351115

Sheheitli, H., and Jirsa, V. K. (2020). A mathematical model of ephaptic interactions in neuronal fiber pathways: Could there be more than transmission along the tracts? *Netw. Neurosci.* 4, 595–610. doi: 10.1162/netn_a_00134

Sherman, S. M., and Guillery, R. W. (1998). On the actions that one nerve cell can have on another: distinguishing "drivers" from "modulators". *Proc. Nat. Acad. Sci.* 95, 7121–7126. doi: 10.1073/pnas.95.12.7121

Sherman, S. M., and Guillery, R. W. (2001). *Exploring the Thalamus*. California: Elsevier.

Sherman, S. M., and Guillery, R. W. (2002). The role of the thalamus in the flow of information to the cortex. *Philosoph. Trans. R. Soc. London.* 357, 1695–1708. doi: 10.1098/rstb.2002.1161

Sherman, S. M., and Guillery, R. W. (2011). Distinct functions for direct and transthalamic corticocortical connections. *J. Neurophysiol.* 106, 1068–1077. doi: 10.1152/jn.00429.2011

Sherrington, C. (1947). *Integrative Action of the Nervous System*. Cambridge: Cambridge University Press.

Sieveritz, B., and Raghavan, R. T. (2021). The central thalamus: gatekeeper or processing hub? *J. Neurosci.* 41, 4954–4956. doi: 10.1523/JNEUROSCI.0573-21.2021

Singer, W. (1999). Neuronal synchrony: a versatile code for the definition of relations? *Neuron* 24, 49–65. doi: 10.1016/S0896-6273(00)80821-1

Sneppen, K., Trusina, A., and Rosvall, M. (2005). Hide-and-seek on complex networks. *EPL (Europhysics Letters)* 69, 853. doi: 10.1209/epl/i2004-10422-0

Spencer, H. (1896). *The Principles of Sociology* (Vol. 6). D. Appleton. doi: 10.5962/bhl.title.61144

Sporns, O. (2012). *Discovering the Human Connectome*. Cambridge, MA: MIT press. doi: 10.7551/mitpress/9266.001.0001

Steriade, M. (2004). Neocortical cell classes are flexible entities. *Nat. Rev. Neurosci.* 5, 121–134. doi: 10.1038/nrn1325

Steriade, M., and Paré, D. (2007). *Gating in Cerebral Networks*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511541735

Sterling, P. (2012). Allostasis: a model of predictive regulation. *Physiol. Behav.* 106, 5–15. doi: 10.1016/j.physbeh.2011.06.004

Sun, J., Taylor, D., and Bollt, E. M. (2015). Causal network inference by optimal causation entropy. *SIAM J. Appl. Dyn. Syst.* 14, 73–106. doi: 10.1137/140956166

Tremblay, S., Testard, C., Inchauspe, J., and Petrides, M. (2022). Non-necessary neural activity in the primate cortex. *bioRxiv*. doi: 10.1101/2022.09.12.506984

Uran, C., Peter, A., Lazar, A., Barnes, W., Klon-Lipok, J., Shapcott, K. A., et al. (2022). Predictive coding of natural images by V1 firing rates and rhythmic synchronization. *Neuron* 110, 1240–1257. doi: 10.1016/j.neuron.2022.01.002

Van den Heuvel, M. P., and Sporns, O. (2011). Rich-club organization of the human connectome. *J. Neurosci.* 31, 15775–15786. doi: 10.1523/JNEUROSCI.3539-11.2011

van der Meij, R., and Voytek, B. (2018). Uncovering neuronal networks defined by consistent between-neuron spike timing from neuronal spike recordings. *Eneuro* 5, 379. doi: 10.1523/ENEURO.0379-17.2018

Varela, F., Lachaux, J. P., Rodriguez, E., and Martinerie, J. (2001). The Brainweb: phase synchronization and large-scale integration. *Nat. Rev. Neurosci.* 2, 229–239. doi: 10.1038/35067550

Varela, F. G., Maturana, H. R., and Uribe, R. (1991). "Autopoiesis: The organization of living systems, its characterization and a model," in *Facets of Systems Science* (Boston, MA: Springer), 559–569. doi: 10.1007/978-1-4899-0718-9_40

Varley, T. F., and Hoel, E. (2022). Emergence as the conversion of information: a unifying theory. *Philosop. Trans. R. Soc. A* 380, 20210150. doi: 10.1098/rsta.2021.0150

Vezoli, J., Vinck, M., Bosman, C. A., Bastos, A. M., Lewis, C. M., Kennedy, H., et al. (2021). Brain rhythms define distinct interaction networks with differential dependence on anatomy. *Neuron* 109, 3862–3878. doi: 10.1016/j.neuron.2021.09.052

Wainwright, M. J. (1999). Visual adaptation as optimal information transmission. *Vision Res.* 39, 3960–3974. doi: 10.1016/S0042-6989(99)00101-7

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440–442. doi: 10.1038/30918

Waxman, S. G. (1972). Regional differentiation of the axon: a review with special reference to the concept of the multiplex neuron. *Brain Res.* 47, 269–288. doi: 10.1016/0006-8993(72)90639-7

Winnubst, J., Bas, E., Ferreira, T. A., Wu, Z., Economo, M. N., Edson, P., et al. (2019). Reconstruction of 1, 000 projection neurons reveals new cell types and organization of long-range con- nectivity in the mouse brain. *Cell* 179, 268–281.e13. doi: 10.1016/j.cell.2019.07.042

Wiskott, L. (2006). How does our visual system achieve shift and size invariance? *Problems Syst. Neurosci.* 26, 322–340. doi: 10.1093/acprof:oso/9780195148220.003.0016

Wolfrum, P. (2010). "Switchyards-Routing Structures in the Brain," in *Information Routing, Correspondence Finding, and Object Recognition in the Brain* (Berlin, Heidelberg: Springer), 69–89. doi: 10.1007/978-3-642-15254-2_4

Womelsdorf, T., Schoffelen, J. M., Oostenveld, R., Singer, W., Desimone, R., Engel, A. K., et al. (2007). Modulation of neuronal interactions through neuronal synchronization. *Science* 316, 1609–1612. doi: 10.1126/science.1139597

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Nat. Acad. Sci.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Zalesky, A., Vu, H. L., Rosberg, Z., Wong, E. W. M., and Zukerman, M. (2007). OBS contention resolution performance. *Perform. Eval.* 64, 357–373. doi: 10.1016/j.peva.2006.06.002

Zeki, S. (2020). "Multiplexing" cells of the visual cortex and the timing enigma of the binding problem. *Eur. J. Neurosci.* 52, 4684–4694. doi: 10.1111/ejn.14921

Zeki, S., and Bartels, A. (1998). The asynchrony of consciousness. *Proc. R. Soc.London.* 265, 1583–1585. doi: 10.1098/rspb.1998.0475

Zhou, S., and Mondragón, R. J. (2004). The rich-club phenomenon in the Internet topology. *IEEE Commun. Lett.* 8, 180–182. doi: 10.1109/LCOMM.2004.823426

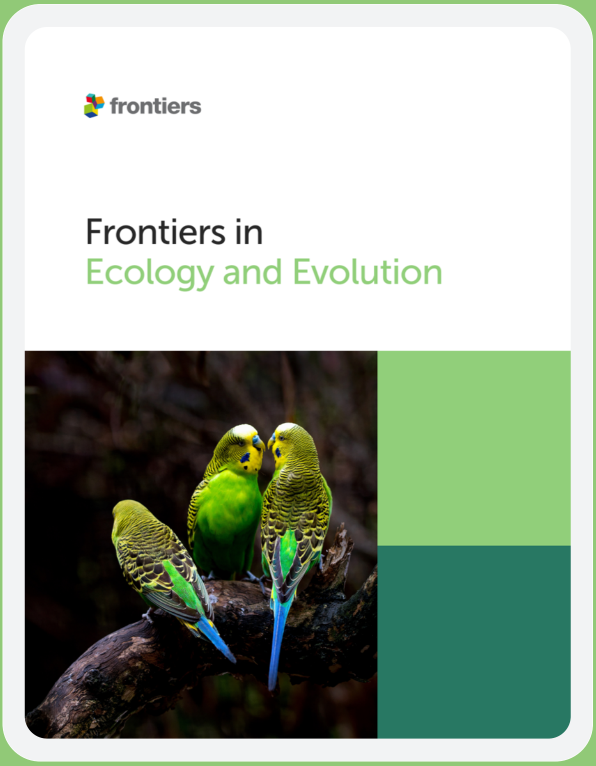
Frontiers in
Ecology and Evolution

Ecological and evolutionary research into our natural and anthropogenic world

This multidisciplinary journal covers the spectrum of ecological and evolutionary inquiry. It provides insights into our natural and anthropogenic world, and how it can best be managed.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact