Studies in Applied Philosophy, Epistemology and Rational Ethics

SAPERE

Vincent C. Müller *Editor*

Philosophy and Theory of Artificial Intelligence 2021

Springer

Studies in Applied Philosophy, Epistemology and Rational Ethics

Volume 63

Studies in Applied Philosophy, Epistemology and Rational Ethics (SAPERE) publishes new developments and advances in all the fields of philosophy, epistemology, and ethics, bringing them together with a cluster of scientific disciplines and technological outcomes: ranging from computer science to life sciences, from economics, law, and education to engineering, logic, and mathematics, from medicine to physics, human sciences, and politics. The series aims at covering all the challenging philosophical and ethical themes of contemporary society, making them appropriately applicable to contemporary theoretical and practical problems, impasses, controversies, and conflicts. Our scientific and technological era has offered "new" topics to all areas of philosophy and ethics– for instance concerning scientific rationality, creativity, human and artificial intelligence, social and folk epistemology, ordinary reasoning, cognitive niches and cultural evolution, ecological crisis, ecologically situated rationality, consciousness, freedom and responsibility, human identity and uniqueness, cooperation, altruism, intersubjectivity and empathy, spirituality, violence. The impact of such topics has been mainly undermined by contemporary cultural settings, whereas they should increase the demand of interdisciplinary applied knowledge and fresh and original understanding. In turn, traditional philosophical and ethical themes have been profoundly affected and transformed as well: they should be further examined as embedded and applied within their scientific and technological environments so to update their received and often old-fashioned disciplinary treatment and appeal. Applying philosophy individuates therefore a new research commitment for the 21st century, focused on the main problems of recent methodological, logical, epistemological, and cognitive aspects of modeling activities employed both in intellectual and scientific discovery, and in technological innovation, including the computational tools intertwined with such practices, to understand them in a wide and integrated perspective. **Studies in Applied Philosophy, Epistemology and Rational Ethics** means to demonstrate the contemporary practical relevance of this novel philosophical approach and thus to provide a home for monographs, lecture notes, selected contributions from specialized conferences and workshops as well as selected Ph.D. theses. The series welcomes contributions from philosophers as well as from scientists, engineers, and intellectuals interested in showing how applying philosophy can increase knowledge about our current world. Initial proposals can be sent to the Editor-in-Chief, Prof. Lorenzo Magnani, lmagnani@unipv.it:

- A short synopsis of the work or the introduction chapter
- The proposed Table of Contents
- The CV of the lead author(s).

For more information, please contact the Editor-in-Chief at lmagnani@unipv.it.

Indexed by SCOPUS, zbMATH, SCImago, DBLP.

All books published in the series are submitted for consideration in Web of Science.

Vincent C. Müller
Editor

Philosophy and Theory of Artificial Intelligence 2021

Springer

Editor
Vincent C. Müller
Friedrich-Alexander University
of Erlangen-Nuremberg (FAU)
Erlangen, Germany

ISSN 2192-6255 ISSN 2192-6263 (electronic)
Studies in Applied Philosophy, Epistemology and Rational Ethics
ISBN 978-3-031-09152-0 ISBN 978-3-031-09153-7 (eBook)
https://doi.org/10.1007/978-3-031-09153-7

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Dedicated to the memory of our colleague and friend
Ivica Crnkovic
(1955–2022)

Advisory Editors

Preface

The papers in this volume result from the 4th conference on the 'Philosophy and Theory of Artificial Intelligence' (PT–AI) that we organised in Gothenburg on 27–28 September, 2021 (see http://www.pt-ai.org/). The local organisation was taken care of by Profs. Ivica Crnkovic and Gordana Dodig-Crnkovic, with support from Chalmers University in Gothenburg.

This conference had been planned for a much earlier date, but several things got in the way, most of all the COVID pandemic. Eventually, we decided that in some way or other we should run PT-AI in 2021, and for most people, this was the first conference where actual physical presence was possible again, though with significant safety measures. It was the first PT-AI conference that was run in a hybrid online/onsite fashion (earlier conferences had already allowed online live listening to keynote talks). As a result of all this, the conference was a bit smaller than usual, with 60 submissions, of which 25 were presented at the meeting. My thanks to the colleagues on the programme committee who worked hard on the double-blind reviewing and assured a very high academic level! The inspiring invited speakers were Virginia Dignum (Umeå, Sweden), Michael Levin (Tufts, USA), David Papineau (KCL, UK), and Shannon Vallor (Edinburgh, UK).

It was very good to see many new faces coming to the field, as well as some established philosophers from other areas showing an interest in the philosophy of AI—which is clearly moving into the mainstream now. The new people and the influences from many different directions are clearly enriching the field and I expect this to continue.

I have to end on a bitter note: In February 2022, our co-organiser Ivica died suddenly. It is extremely sad to see such a rich life cut short, and many people left behind with a huge gap in their lives—first of all, his wife and his children. At the same time, I am also grateful that I had the chance to know Ivica and learn from him, academically and as a human being.

Eindhoven, The Netherlands
April 2022

Vincent C. Müller
http://www.sophia.de/

Contents

Theory

Cognitive Architectures Based on Natural Info-Computation 3
Gordana Dodig-Crnkovic

Artificial Intelligence Systems, Responsibility and Agential Self-Awareness 15
Lydia Farina

Models, Algorithms, and the Subjects of Transparency 27
Hajo Greif

Autonomy of Attention 39
Kaisa Kärki

Toward Out-of-Distribution Generalization Through Inductive Biases 57
Caterina Moruzzi

Is There a Trade-Off Between Human Autonomy and the 'Autonomy' of AI Systems? 67
Carina Prunkl

Towards a Taxonomy for the Opacity of AI Systems 73
Alessandro Facchini and Alberto Termine

Validating Non-trivial Semantic Properties of Autonomous Robots 91
Jiří Wiedermann and Jan van Leeuwen

Ethics

Dignity in Digital Ethics 107
Marcel Becker

An Enactive Approach to Value Alignment in Artificial Intelligence: A Matter of Relevance 119
Michael Cannon

The Problem of AI Influence 137
Laura Crompton

Artificial General Intelligence and the Common Sense Argument 155
Olle Häggström

Moral Status of AI Systems: Evaluation of the Genetic Account 161
Leonhard Kerkeling

Deception by Default 171
András Kornai

Robot Rights in Joint Action 179
Guido Löhr

Is It Likely that We Are Living in a Computer Simulation? 193
Ralf Stapelfeldt

Human-AI Friendship: Rejecting the Appropriate Sentimentality Criterion 209
Dan Weijers and Nick Munn

AI Risk Skepticism 225
Roman V. Yampolskiy

Theory

Cognitive Architectures Based on Natural Info-Computation

Gordana Dodig-Crnkovic

Abstract At the time when the first models of cognitive architectures have been proposed, some forty years ago, understanding of cognition, embodiment and evolution was substantially different from today's. So was the state of the art of information physics, information chemistry, bioinformatics, neuroinformatics, computational neuroscience, complexity theory, self-organization, theory of evolution, as well as the basic concepts of information and computation. Novel developments support a constructive interdisciplinary framework for cognitive architectures based on natural morphological computing, where interactions between constituents at different levels of organization of matter-energy and their corresponding time-dependent dynamics, lead to complexification of agency and increased cognitive capacities of living organisms that unfold through evolution. Proposed info-computational framework for naturalizing cognition considers present updates (generalizations) of the concepts of *information, computation*, *cognition,* and *evolution* in order to attain an alignment with the current state of the art in corresponding research fields. Some important open questions are suggested for future research with implications for further development of cognitive and intelligent technologies.

Keywords Cognitive architectures · Natural information · Natural computation · Naturalized cognition · Cognitive evolution · Extended evolutionary synthesis · Basal cognition

G. Dodig-Crnkovic (✉)
Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden
e-mail: dodig@chalmers.se

Division of Computer Science and Software Engineering, School of Innovation, Design and Engineering, Mälardalen University, Västerås, Sweden

V. C. Müller (ed.), *Philosophy and Theory of Artificial Intelligence 2021*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 63,
https://doi.org/10.1007/978-3-031-09153-7_1

1 Introduction

In 1958 John von Neumann wrote "The computer and the brain" (von Neumann, 1958)—the book describing information processing architecture of computers as based on then-current understanding of brain organization, with separate memory, input/output unit, arithmetic/logic unit, and a control unit. Von Neumann architecture is still in use. However, understanding of the brain have changed radically (Gazzaniga et al., 2019; Damasio, 2021), as well as the possibilities of distributed concurrent and intrinsic natural computing (Crutchfield et al., 2010; Burgin & Dodig-Crnkovic, 2015).

We may hope that new understanding of the brain and cognition as well as computation possibilities (information processing, structures, and dynamics) will bring about new nature-inspired (biomimetic) cognitive computational architectures. One development in that direction is neuromorphic computing, inspired by human brain function. Compared to von Neumann architectures, it puts very different requirements on the cognitive computational system, such as: the use of the same elements for processing and memory/storage of information; variation of electrical properties according to the Hebbian learning (electronic synapses), auto-oscillation generation mode, stable chains of signal transfer, and capacity of self-organization into 3D systems for the materials used for electronic compounds, all of which is mimicking intrinsic brain functions (Erokhin, 2022).

A recent overview of 40 years of research and practical applications in cognitive architectures, (Kotseruba & Tsotsos, 2020), addresses the adequacy of existing cognitive architectures in modelling of the *core cognitive abilities in humans*, including perception, attention, action, memory, learning, and reasoning. Apart from presenting the state-of-the-art of the research through 84 human-level cognitive architectures, authors briefly mention deep learning, and why it does not qualify as *unified model of cognition in humans.* We will come back to the relation between recent developments in deep learning and understanding of (human) cognitive processes. Interesting recent work (Stocco et al., 2021) presents an analysis of the *human connectome data* that supports the notion of a "*Common Model of Cognition*" for *human and human-like* intelligence across multiple brain regions and cognitive domains. However, our focus is not on human and human-like cognition, but on the evolutionary origins of cognition and its development from basal cognition to the diversity of forms of cognition in all living organisms.

The present account introduces natural computational cognitive architectures, not included in Kotseruba & Tsotsos (2020), in the first place because they do not address exclusively human level cognition, but treats all living beings as cognizing agents. In this naturalistic approach, the underlying assumption is that cognition in nature is a manifestation of biological processes (that subsume chemical and physical levels) in all living organisms *from single cells to humans* (Dodig-Crnkovic, 2007; Jagers op Akkerhuis, 2010; Lyon, 2005, 2015a; Lyon & Kuchling, 2021; Maturana & Varela, 1992; Stewart, 1996).

Recently Piccinini (2020) made a step beyond the usual assumption that cognition (and intelligence) necessarily presuppose human agent. Piccinini addresses biological cognition in any organism with nervous systems as a result of neurocomputation. This approach does not go the full way to include all living organisms, even those without nervous systems, in spite of new findings of biologists that "*cognitive operations we usually ascribe to brains—sensing, information processing, memory, valence, decision making, learning, anticipation, problem solving, generalization and goal directedness—are all observed in living forms that don't have brains or even neurons*" (Levin et al., 2021). Similar arguments for biogenic nature of cognition can be found in Lyon (2015b), Lyon et al. (2021).

Grounded in the empirical and theoretical work on cognition and its evolution in nature (Walker et al., 2017; Dodig-Crnkovic, 2017a), from basal/basic/primitive/elementary/cellular, to complex form of human cognition (Dodig-Crnkovic, 2014, 2020; Levin et al., 2021; Lyon et al., 2021; Manicka & Levin, 2019; Stewart, 1996), with natural information processing (natural computation) as a basis, info-computational approach can be used to identify several topics in the research of cognition that need more study.

First of all, in order to understand cognition, we must put it in the context of process of evolution (Dobzhansky, 1973). The process of evolution of nervous system and brain, as well as sensory organs which are central for human cognition, deserve special attention.

Lyon with collaborators propose "reframing cognition by getting down to biological basics" in an article that is part of the Philosophical Transactions of the Royal Society B theme issue 'Basal cognition: conceptual tools and the view from the single cell' which explores in depth the cognition on the single-cellular level in its evolutionary context (Lyon et al., 2021).

As a contribution to the attempt at bridging the gap between high level human cognition and the unicellular basal cognition, it is instructive to study intermediate steps. Recently we could read that "Brainless sponges contain early echoes of a nervous system", as described in Science News. In sponges we can trace back the origin of capacities of "higher order" cognition, resembling those existing in the human nervous system, which point to the evolution of nervous cells from the ordinary somatic cells of simple organisms. This recent discovery of "neuroid cells" in sponges attracted a lot of publicity showing that some cells evolved ability and specialized in connecting inside (digestive system) with the outside (source of food), having genes in common with neurons, and playing similar role for simple and complex organisms. Neuroid ("proto-neural") cells are in contact with cellular cilia, a short microscopic hairlike structure on the surface of cells, either causing currents in the surrounding fluid, or, in some protozoans, providing propulsion, according to https://www.dictionary.com/browse/cilia. Signals from neuroid cells prompt cilia to start or stop waving, and thus control feeding (Pennisi, 2021).

The rest of the paper is organized as follows. Section 2 presents naturalized cognition and human thinking, fast and slow, while Sect. 3 addresses the open questions of cognitive architectures and natural info-computation. Section 4 offers conclusions.

2 Naturalized Cognition and Human Thinking, Fast and Slow

Thus the organic body of each living being is a kind of divine machine or natural automaton which infinitely surpasses all artificial automata. For a machine made by the skill of man is not a machine in each of its parts... But the machines of nature, namely, living bodies, are still machines in their smallest parts ad infinitum. It is this that constitutes the difference between nature and art, that is to say, between the divine art and ours.

(Leibniz, 1898) Monadologie §64, p. 254.

Cognitive architectures started as a research field with the goal to model *human mind* and build *human-level artificial intelligence*. By connecting models and mechanisms with observed cognitive/intelligent behaviors, they contributed to cognitive science and AI. However, cognition in nature appears throughout biological systems (Almér et al., 2015; Baluška & Levin, 2016; Lyon, 2005, 2015a; Lyon et al., 2021) and it is important to understand its evolutionary development from the basal/basic/elementary cognition to the human level cognition (Levin et al., 2021; Manicka & Levin, 2019).

This naturalized evolutionary approach to cognition is based on the view of hierarchical recursive structure of information processing in nature, in living organisms from cells, to tissues, organs, organisms and their groups—all of them communicating at different levels of organization by exchanging specific types of information—physical (elementary particles, electro-magnetic), chemical (electric, molecular), biological, and symbolic (signs, languages).

In humans, two basic *functional abstractions* of cognitive system have been proposed, System 1 (reflexive, non-conscious, automatic, intuitive information processing, which is fast) and System 2 (reflective, conscious, reasoning and decision making, which is slow) (Kahneman, 2011; Tjøstheim et al., 2020).

Within AI, the field of artificial neural networks with deep learning have made an impressive progress in modeling perception on the level of data/signal processing. Deep learning level corresponds to Kahneman's fast, intuitive System 1. Current developments in AI (addressing *human-level cognition*) are continuing towards modelling System 2, symbolic reasoning (Russin et al., 2020).

It has long been recognized that mechanisms of cognition based on natural computation are far more sophisticated than the machine-like classical computationalist models based on abstract symbol manipulation (Kampis, 1991). They conform to the view expressed by Witzany & Baluška (2012) that *rule-based machines are not good enough models of natural cognition.* Compare to the Leibniz insight from the quote above.

Natural/physical/intrinsic/morphological/computation presupposes embodiment of information processing. *Embodiment is the fundamental feature of cognition*, which implies that *valence, affect, feelings and emotions* must be taken into account as constitutive elements in the models of cognition (Damasio, 1999; Dodig-Crnkovic, 2017a; Dodig-Crnkovic & Giovagnoli, 2017; Lyon & Kuchling, 2021; Watanabe et al., 2017). They affect both System 1 and System 2 information processing.

3 Open Questions of Cognitive Architectures and Natural Info-Computation

With the present development in scientific research as well as cognitive and intelligent computing it is becoming important to update computational approaches to cognitive architectures. Currently, there are several interesting open questions worth more exploration.

3.1 Biomimetic Design of Cognitive Architectures. What Is "Biologically Plausible"?

Proposals to learn from nature about cognition are old (Maturana & Varela, 1992; Stewart, 1996; Lyon, 2005, 2015a; Dodig-Crnkovic, 2007; Jagers op Akkerhuis, 2010; Lyon & Kuchling, 2021), but they have recently gained a lot of prominence in the form of biomimetic design (Joyee et al., 2020). Can our newly acquired insights into cognition on different levels of organization in nature be applied to improve cognitive architectures?

Russin et al. (2020) argue that deep learning (corresponding to Kahneman's System 1) needs an equivalent of "prefrontal cortex" that would play the role of System 2 (slow, reflective information processes). This is in agreement with Marblestone et al. (2016), who suggest increased integration of deep learning and neurosciences. Similar ideas are put forward by Dodig-Crnkovic (2020) with the argument that natural morphological computation should be used to study function of meta-learning (learning to learn) in humans (function of prefrontal cortex), other living organisms, and intelligent machines.

Here we should recognize that Bengio's and Kahneman's interpretations of System 1 and System 2 are not identical, which was evident from the discussion at AAAI-2020 conference, Fireside Chat with Lecun, Hinton, Bengio and Kahneman https://vimeo.com/390814190. However, the details of interpretations are not essential for our current exposition.

To this current discussion about how Bengio-Lecun-Hinton's interpretation relates to Kahneman's views, one should add critical voices questioning dual-process theories in general as inadequate, as presented in Osman (2004) review. Osman proposes replacing the dualist (dual-aspect) approach with "a single-system framework that conjectures that different types of reasoning arise through the graded properties of the representations that are utilized while reasoning and the different functional roles that consciousness has in cognition", arguing for the framework, unifying the different forms of reasoning, identified by dual-process theorists, under a single system. Bengio-Lecun-Hinton's interpretation seems to be closer to Osman who searches for connections between System 1 and System 2, especially Bengio elucidates the role of consciousness for learning (Russin et al., 2020).

3.2 Cognitive Behaviors and Their Simulation, Emulation and Engineering

In the special report "Can We Copy the Brain?" (The Editors of IEEE Spectrum, 2017), the founder of the Blue Brain Project, Henry Markram discusses complexities of the brain and necessity of learning about the details of its functioning on different levels of organization. He also discusses possibility to simulate the brain with molecular and cellular level simplified and encapsulated. There are two open questions that run in parallel, providing an opportunity for two-way learning between computing and neuroscience (Rozenberg & Kari, 2008). The questions are: first, how cognition works and develops in nature, and second, how we can model, simulate, emulate and engineer it in computational artifacts.

Work of Michael Levin (https://ase.tufts.edu/biology/labs/levin) suggests broad range of applications for nature-inspired cognitive architectures based on biological cognition connecting genetic networks, cytoskeleton, neural networks, tissue/organ, and organism with the group (social) levels of information processing. Levin shows how biology has been computing through somatic memory (information storage) and biocomputation/decision making in pre-neural bioelectric networks, before the development of neurons and brains. Fields et al. (2020) summarize:

> *Importantly, neurons utilize ancient mechanisms such as ion channels, electrical synapses, and neurotransmitters that also operate throughout the body in many non-excitable tissues and predate the evolution of specialized neurons. We here propose a model in which both neuronal signals and non-neural bioelectric patterning signals arise from modifications of conserved basic machinery, and co-evolved to function to control both organismal behavior and development.*

Insights from biocognition can help the development of new AI platforms, applications in targeted drug delivery, regenerative medicine and cancer therapy, nano-technology, synthetic biology, artificial life, and much more.

3.3 Computational Efficiency of Natural Computing

Computational efficiency and performance are important features, often left outside when discussing computational models of cognition. However, with the increased ubiquity of computing, this aspect becomes essential. Natural cognitive computing, being particularly resource effective, can provide ideas for future developments towards more resource-efficient computational architectures (Usman et al., 2019; Nature Editorial, 2019).

The question of computational efficiency has also been addressed by biomimetic neuromorphic computing which is mimicking the neural structure and functions of the human brain, together with probabilistic computing, with algorithmic approaches to the uncertainty, ambiguity, and contradiction in nature (Ackerman, 2019). More

learning from nature about computational efficiency is needed that will inform biomimetic designs of cognitive architectures.

3.4 Time Aspect of Cognitive Models of Naturalized Cognition

Cognitive models today take the mind/brain to be reactive, with information processing starting with a stimulus and ending with a response (Bechtel, 2013). However, cells are inherently active, neurons are sustained oscillators, exhibiting electrochemical oscillations even in the absence of stimuli. Input data/information presents stimuli that *modulate existing endogenous oscillations* (Bechtel, 2013). In the book "Rhythms of the Brain" Buzsáki (2009) describes the important role that spontaneous activity of neurons plays. Spontaneous firing of neurons is the very basis of human cognition when it comes to its time aspects. A self-organized timing of oscillations has co-evolved as the main organizational principle of neuronal activity. Global computation (on multiple spatial and temporal scales) is enabled by small-world-connectivity of neurons in the cerebral cortex. In a small-world setting, any two of nodes are connected through a short sequence of intermediary nodes. Cortical system is in a metastable state, synchronized through weak links between network oscillations in constant interactions. Oscillator frequency determines periods of receiving and transferring information.

Based on studies of oscillations, neural computations and learning, Penagos et al. (2017) propose that "*precisely coordinated representations across brain regions allow the inference and evaluation of causal relationships to train an internal generative model of the world*." Training starts while awake, and processing continues during sleep when periodic nested oscillations induce hierarchical processing of information. Authors suggest that "general inference, prediction and insight" supporting an internal model for generalization and adaptive behavior is enabled through periodic states of sleep.

Related is the synaptic plasticity of the brain which changes its connections through the long-term potentiation (Hebbian and non-Hebbian), considered to be a basis for learning and memory. Oscillatory behavior is not only characteristics of the human brain. Similar oscillatory rhythms have been observed in the brains of mice. Being made of *oscillators, biological neural networks are able to filter inputs and to resonate with noise.* Unlike those observed oscillatory time behaviors in the biological brains, that appear as a result of their physical embodiment, artificial neural networks have no such temporal coupling and synchronizing mechanisms. It is an open question how essential this oscillatory behavior and metastability are for "fine tuning to the world" and if their function can be obtained in a different way.

On the global level of unified theories of cognition, time aspect (Anderson, 2002) manifests itself in terms of Newell's *bands of cognition* (Newell, 1994)—the biological "10 ms band", cognitive, rational, and social ("long-term") bands. How important

is it to have all of them represented and how detailed? Here we talk about understanding of temporal aspects of cognition as organized hierarchically in a metastable state, constantly tuning to the environment. Coordination obtained through communication is central for connecting different levels, from molecules to thoughts, in the same coordination dynamics (Kelso et al., 2013). Through the interplay with the environment this process results in *eigenstates* (Foerster, 2003). Technological approaches to cognitive models of brain-like computer, based on frequency-fractal computing are proposed by Ghosh et al. (2014), Singh et al. (2020). In short, time aspect of cognitive models of naturalized cognition deserves more attention.

4 Conclusion

Modelling cognitive processes as natural computation/physical computation/morphological computation (natural information processing), we can better understand cognition as it evolves in living beings (Dodig-Crnkovic, 2017a).

Identified open questions deserving further attention include biomimetic design of cognitive architectures and meaning of the expectation of biological plausibility for understanding of cognition and for technological applications; cognitive behaviors in nature and their simulation, emulation, and engineering; taking advantage of computational efficiency of natural computing, and deeper understanding of time aspects of cognition on hierarchy of levels of organization and in evolutionary context.

The info-computational framework considers state of the art in the research and applications of information and computation, as well as relevant parts of information physics, information chemistry, bioinformatics, neuroinformatics, computational neuroscience, complexity theory, self-organization, and the developments in the theory of evolution, for naturalizing cognition. It requires generalization of several fundamental concepts, as follows:

Information is seen as the fabric of the universe/nature. For a cognitive agent, information is the basis or reality on which behavior is based.

Computation is *information processing (dynamics of information).*

Cognition is characteristics of all living forms, not only humans or organisms with nervous systems. Cognition is a network of life-sustaining processes that enables every living organism to perceive its environment, react adequately and adapt so to survive as individuals and species.

Evolution is understood as *Extended evolutionary synthesis*, which considers that not only random mutations, but also sequences of changes caused by laws of physics and chemistry (that can be described as *morphological computation*) leads to the development of new structures which are then exposed to natural selection (Jablonka & Lamb, 2014; Laland et al., 2015).

Parallel development of our understanding of cognition as natural phenomenon and its technological implementations inform each other in a recursive manner (Rozenberg & Kari, 2008; Bondgard & Levin, 2021). As we have seen, learning from

nature and biomimetic design necessitate interdisciplinary approaches to computing as exemplified in approaches in Dodig-Crnkovic (2017b), also argued for in Esposito et al. (2018).

Development towards biomimetic architectural design inspired by natural intrinsic morphological computing promises new resource-effective cognitive architectures on different levels of complexity—from basal cognition that can be used in nanotechnology to complex cognition needed for social robotics.

References

Ackerman, E. (2019). Intel labs director talks quantum, probabilistic, and neuromorphic computing - IEEE spectrum. *IEEE Spectrum*.

Almér, A., Dodig-Crnkovic, G., & von Haugwitz, R. (2015). Collective cognition and distributed information processing from bacteria to humans. In *Proc. AISB Conference Kent, April 2015*.

Anderson, J. R. (2002). Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science*. https://doi.org/10.1207/s15516709cog2601_3

Baluška, F., & Levin, M. (2016). On having no head: Cognition throughout biological systems. *Frontiers in Psychology, 7*, 902.

Bechtel, W. (2013). The endogenously active brain: The need for an alternative cognitive architecture. *Philosophia Scientiæ, 17*(2), 3–30.

Bondgard, J., & Levin, M. (2021). Living things are not (20th century) machines: Updating mechanism metaphors in light of the modern science of machine behavior. *Frontiers in Ecology and Evolution, 9*, 147.

Burgin, M., & Dodig-Crnkovic, G. (2015). A taxonomy of computation and information architecture. In M. Galster (Ed.), *Proceedings of the 2015 European Conference on Software Architecture Workshops (ECSAW '15)*. ACM Press. https://doi.org/10.1145/2797433.2797440

Buzsáki, G. (2009). *Rhythms of the brain*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195301069.001.0001

Crutchfield, J., Ditto, W., & Sinha, S. (2010). Introduction to focus issue: Intrinsic and designed computation: Information processing in dynamical systems—Beyond the digital hegemony. *Chaos*, *20*, 037101.

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. Harcourt Brace and Co.

Damasio, A. R. (2021). Feeling & knowing: Making minds conscious. *Cognitive Neuroscience, 12*(2), 65–66.

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *American Biology Teacher*, *35*(3).

Dodig-Crnkovic, G. (2007). Epistemology naturalized: The info-computationalist approach. *APA Newsletter on Philosophy and Computers, 06*(2), 9–13.

Dodig-Crnkovic, G. (2014). Modeling life as cognitive info-computation. In A. Beckmann, E. Csuhaj-Varjú, & K. Meer (Eds.), *Computability in Europe 2014. LNCS* (pp. 153–162). Springer. http://arxiv.org/abs/1401.7191

Dodig-Crnkovic, G. (2017a). Nature as a network of morphological infocomputational processes for cognitive agents. *European Physical Journal, 226*, 181–195. https://doi.org/10.1140/epjst/e2016-60362-9

Dodig-Crnkovic, G. (2017b). Nature as a network of morphological infocomputational processes for cognitive agents. *European Physical Journal: Special Topics*, *226*(2). https://doi.org/10.1140/epjst/e2016-60362-9

Dodig-Crnkovic, G. (2020). Natural morphological computation as foundation of learning to learn in humans, other living organisms, and intelligent machines. *Philosophies*. https://doi.org/10.3390/philosophies5030017

Dodig-Crnkovic, G., & Giovagnoli, R. (2017). *Representation and reality in humans, other living organisms and intelligent machines* (G. Dodig-Crnkovic & R. Giovagnoli, Eds.). Springer International Publishing. https://doi.org/10.1007/978-3-319-43784-2

Erokhin, V. (2022). *Fundamentals of organic neuromorphic systems*. Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-79492-7

Esposito, A., Faundez-Zanuy, M., Morabito, F. C., & Pasero, E. (2018). *Multidisciplinary approaches to neural computing*. Springer International Publishing.

Fields, C., Bischof, J., & Levin, M. (2020). *Morphological coordination: A common ancestral function unifying neural and non-neural signaling, 35*, 16–30.

Foerster, H. von. (2003). *Understanding understanding: Essays on cybernetics and cognition.* Springer Berlin Heidelberg.

Gazzaniga, M., Ivry, R. B., & Mangun, G. R. (2019). *Cognitive neuroscience: The biology of the mind* (5th ed.). WW Norton & Company.

Ghosh, S., Aswani, K., Singh, S., Sahu, S., Fujita, D., & Bandyopadhyay, A. (2014). Design and construction of a brain-like computer: A new class of frequency-fractal computing using wireless communication in a supramolecular organic, inorganic system. *Information, 5*(1), 28–100. https://doi.org/10.3390/info5010028

Jablonka, E., & Lamb, M. (2014). *Evolution in four dimensions: Genetic, epigenetic, behavioral, and symbolic variation in the history of life* (Revised ed.). Life and mind: Philosophical issues in biology and psychology. A Bradford Book. MIT Press.

Jagers op Akkerhuis, G. (2010). *The operator hierarchy: A chain of closures linking matter, life and artifi cial intelligence*. Ph.D. dissertation, Radboud University Nijmegen.

Joyee, E. B., Szmelter, A., Eddington, D., & Pan, Y. (2020). 3D printed biomimetic soft robot with multimodal locomotion and multifunctionality. *Soft Robotics*. https://doi.org/10.1089/soro.2020.0004

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kampis, G. (1991). *Self-modifying systems in biology and cognitive science: A new framework for dynamics, information, and complexity*. Pergamon Press.

Kelso, S. J. A., Dumas, G., & Tognoli, E. (2013). Outline of a general theory of behavior and brain coordination. *Neural Networks, 37*, 120–131.

Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-018-9646-y

Laland, K. N., Uller, T., Feldman, M. W., Sterelny, K., Müller, G. B., Moczek, A., et al. (2015). The extended evolutionary synthesis: Its structure, assumptions and predictions. *Proceedings of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rspb.2015.1019

Leibniz, G. W. (1898). *Monadology and other philosophical writings* (Robert Latta, Ed.). Clarendon Press/Oxford University Press.

Levin, M., Keijzer, F., Lyon, P., & Arendt, D. (2021). Uncovering cognitive similarities and differences, conservation and innovation. *Philosophical Transactions of the Royal Society B, 376*, 20200458.

Lyon, P. (2005). The biogenic approach to cognition. *Cognitive Processing, 7*, 11–29.

Lyon, P. (2015a). The cognitive cell: Bacterial behaviour reconsidered. *Frontiers in Microbiology, 6*, 264.

Lyon, P., Keijzer, F., Arendt, D., & Levin, M. (2021). Reframing cognition: Getting down to biological basics. *Philosophical Transactions of the Royal Society B, 376*, 20190750.

Lyon, P., & Kuchling, F. (2021). Valuing what happens: A biogenic approach to valence and (potentially) affect. *Philosophical Transactions of the Royal Society B, 376*, 2019075220190752.

Manicka, S., & Levin, M. (2019). The cognitive lens: A primer on conceptual tools for analysing information processing in developmental and regenerative morphogenesis. *Philosophical Transactions of the Royal Society B*, *374*(1774).

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*. https://doi.org/10.3389/fncom.2016.00094

Maturana, H., & Varela, F. (1992). *The tree of knowledge*. Shambala.

Nature Editorial. (2019). How to make computing more sustainable. *Nature, 573*, 310.

Newell, A. (1994). *Unified theories of cognition* (Reprint ed.). Harvard University Press.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review, 11*(6), 988–1010. https://doi.org/10.3758/BF03196730

Penagos, H., Varela, C., & Wilson, M. A. (2017). Oscillations, neural computations and learning during wake and sleep. *Current Opinion in Neurobiology*. https://doi.org/10.1016/j.conb.2017.05.009

Pennisi, E. (2021). Sponge innards suggest how nerve cells evolved. *Science*. https://doi.org/10.1126/science.acx9579

Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford scholarship online.

Rozenberg, G., & Kari, L. (2008). The many facets of natural computing. *Communications of the ACM, 51*, 72–83.

Russin, J., O'Reilly, R. C., & Bengio, Y. (2020). Deep learning needs a prefrontal cortex. *Workshop "Bridging AI and Cognitive Science" (ICLR 2020)*.

Singh, P., Saxena, K., Singhania, A., Sahoo, P., Ghosh, S., Chhajed, R., et al. (2020). A self-operating time crystal model of the human brain: Can we replace entire brain hardware with a 3D fractal architecture of clocks alone? *Information, 11*(5), 238.

Stewart, J. (1996). Cognition = life: Implications for higher-level cognition. *Behavioral Processes, 35*, 311–326.

Stocco, A., Sibert, C., Steine-Hanson, Z., Koh, N., Laird, J. E., Lebiere, C. J., & Rosenbloom, P. (2021). Analysis of the human connectome data supports the notion of a "common model of cognition" for human and human-like intelligence across domains. *NeuroImage*, *235*, 118035. https://doi.org/10.1016/j.neuroimage.2021.118035

The Editors of IEEE Spectrum. (2017). Special report: Can we copy the brain? *IEEE Spectrum*. https://spectrum.ieee.org/static/special-report-can-we-copy-the-brain

Tjøstheim, T. A., Stephens, A., Anikin, A., & Schwaninger, A. (2020). The cognitive philosophy of communication. *Philosophies*. https://doi.org/10.3390/philosophies5040039

Usman, M. J., Ismail, A. S., Abdul-Salaam, G., Chizari, H., Kaiwartya, O., Gital, A. Y., et al. (2019). Energy-efficient nature-inspired techniques in cloud computing datacenters. *Telecommunication Systems, 71*, 275–302.

von Neumann, J. (1958). *The computer and the brain*. Yale Univ Press.

Walker, S. I., Davies, P., & Ellis, G. (2017). *From matter to life information and causality*. Cambridge University Press. Kindle Edition.

Watanabe, S., Hofman, M. A., & Toru, S. (Eds.). (2017). *Evolution of the brain, cognition, and emotion in vertebrates*. Springer.

Witzany, G., & Baluška, F. (2012). Turing: A formal clash of codes. *Nature, 483*, 541.

Artificial Intelligence Systems, Responsibility and Agential Self-Awareness

Lydia Farina

Abstract This paper investigates the claim that artificial Intelligence Systems cannot be held morally responsible because they do not have an ability for agential self-awareness e.g. they cannot be aware that they are the agents of an action. The main suggestion is that if agential self-awareness and related first person representations presuppose an awareness of a self, the possibility of responsible artificial intelligence systems cannot be evaluated independently of research conducted on the nature of the self. Focusing on a specific account of the self from the phenomenological tradition, this paper suggests that a minimal necessary condition that artificial intelligence systems must satisfy so that they have a capability for self-awareness, is having a minimal self defined as 'a sense of ownership'. As this sense of ownership is usually associated with having a living body, one suggestion is that artificial intelligence systems must have similar living bodies so they can have a sense of self. Discussing cases of robotic animals as examples of the possibility of artificial intelligence systems having a sense of self, the paper concludes that the possibility of artificial intelligence systems having a 'sense of ownership' or a sense of self may be a necessary condition for having responsibility.

Keywords AI responsibility · Artificial self · Agential self awareness · Personal identity

1 Introduction

The current debate on the possibility of attributing moral responsibility to artificial Intelligence Systems focuses on the concepts of autonomy or consciousness (Coeckelbergh, 2020; Müller, 2020). According to some views, moral responsibility for artificial intelligence systems is considered impossible because they lack either autonomy or consciousness and these are necessary (but not sufficient) for moral

L. Farina (✉)
Department of Philosophy, University of Nottingham, Nottingham NG7 2RD, UK
e-mail: lydia.farina@nottingham.ac.uk

V. C. Müller (ed.), *Philosophy and Theory of Artificial Intelligence 2021*,
Studies in Applied Philosophy, Epistemology and Rational Ethics 63,
https://doi.org/10.1007/978-3-031-09153-7_2