Philosophical Problems in Science

Zagadnienia Filozoficzne w Nauce

ISSN 0867-8286 (print format)

e-ISSN 2451-0602 (electronic format)

Editorial Office
Zagadnienia Filozoficzne w Nauce
Wydział Filozoficzny UPJPII
ul. Kanonicza 9, 31-002 Kraków
POLAND

e-mail: info@zfn.edu.pl
www.zfn.edu.pl

Publisher: Copernicus Center Press Sp. z o.o.
pl. Szczepański 8, 31-011 Kraków POLAND
tel. (+48) 12 448 14 12
e-mail: marketing@ccpress.pl
www.ccpress.pl

Philosophical Problems in Science

Zagadnienia Filozoficzne w Nauce

LXXIII ■ 2022

Editorial

Od redakcji

Roman Krzanowski
Editorial note 7

Articles

Artykuły

Gordana Dodig-Crnkovic
In search of a common, information-processing, agency-based framework for anthropogenic, biogenic, and abiotic cognition and intelligence 17

Javier Toscano
But seriously: what do algorithms want? Implying collective intentionalities in algorithmic relays—a distributed cognition approach 47

Alice Martin, Mathieu Magnaudet, Stéphane Conversy
Modelling interactive computing systems: Do we have a good theory of what computers are? 77

Hyungrae Noh
Shannon-inspired information in the clinical use of neural signals concerning post-comatose patients 121

Łukasz Mścisławski
Is information ontological or physical, or is it perhaps something else? Some remarks on Krzanowski's approach to the concept of information 147

Kristina Šekrst, Sandro Skansi
Machine learning and essentialism 171

Roman Krzanowski, Pawel Polak
The meta-ontology of AI systems with human-level intelligence 197

Krzysztof Sołoducha
Analysis of the implications of the Moral Machine project as an implementation of the concept of coherent extrapolated volition for building clustered trust in autonomous machines 231

Essays
Eseje

Adam Olszewski
Will a human always outsmart a computer? An essay 259

Kazimierz Trzęsicki
Perspective on Turing paradigm: An essay 281

Review articles
Artykuły recenzyjne

Paweł Polak
Beyond epistemic concepts of information: The case of ontological information as philosophy in science 335

Paweł Polak
Why is neuron modeling of particular philosophical interest? 347

Łukasz Mścisławski
Is AI case that is explainable, intelligible or hopeless? 357

Editorial note

Roman Krzanowski

This special edition of Philosophical Problems in Science (*Zagadnienia Filozoficzne w Nauce* or ZFN) focuses on concepts of information and computing. On reading this issue, you may be surprised by the absence of traditional perspectives and themes that one would usually expect from such collections, but this apparent oversight is deliberate. The eight papers collected in this special edition of ZFN bring together perspectives that aim to inspire readers rather than confirm concepts that have already been researched. The main motivation behind this collection is a desire to explore the philosophical dimensions of computing and information sciences. Thus, for anyone looking for new ideas related to the philosophy of computing and information and wondering what is on the horizon, this special edition of ZFN may be the place to start.

The collection begins with a paper by Gordana Dodig-Crnkovic (Chalmers University of Technology) entitled "In Search of a Common, Information-processing, Agency-based Framework for Anthropogenic, Biogenic, and Abiotic Cognition and Intelligence." This paper aims to provide a general introduction to advances in natural computing and information processing in order to:

> better to understand mechanisms of cognition and intelligence as they appear in nature. New understandings of information and processes of physical (morphological) computation contribute to novel possibilities that can be used to inspire the development of abiotic cognitive systems (cognitive robotics), cognitive computing and artificial intelligence.

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 7–13

This paper also includes extensive, up-to-date references that will help those wishing to explore this topic further by serving as a guide to the state of current research in natural computing.

Next comes a paper by Javier Toscano (Center for Advanced Internet Studies) entitled "But Seriously: What Do Algorithms Want? Implying Collective Internationalities in Algorithmic Relays—a Distributed Cognition Approach." This paper presents the concept of algorithms not as it is usually conceived, namely as a sequence of logical steps, but rather as a:

> larger construct that draws upon sociological and anthropological theories that underline social practices to propose expanding our understanding of an algorithm through the notion of "collective internationalities."

This paper contributes to the discussion about the role that "intentionalities play in understanding socio-structured practices and cognitive ecologies." Furthermore, an extensive bibliography offers up-to-date sources on the paper's topic.

Another viewpoint for the fundamental concepts of computing is put forward by Alice Martin, Mathieu Magnaudet, and Stéphane Conversy (Interactive Informatics Team of ENAC Research Lab) in a paper entitled "Modelling Interactive Computing Systems: Do We Have a Good Theory of What Computers Are?" This paper discusses the conceptualization of interactive computer systems. According to the authors, this type of computing does not receive enough attention from philosophers and computer scientists, so their paper attempts to fill this gap. The paper surveys three areas in which interaction models can be framed: works on concurrency by Milner, works on reactive Turing machines, and works on interaction as a new computing paradigm. For each of these models, the authors present the motivation behind it, summarize its accounting of interaction and its

legacy, and point out issues related to our understanding of computers. The provided references also provide a detailed review of the available literature for this topic.

The interdisciplinary approach to the philosophy of information is a key topic of the next paper by Hyungrae Noh (Sunchon National University), which is titled "Shannon-Inspired Information in the Clinical Use of Neural Signals Concerning Post-Comatose Patients." This paper links the two domains of medicine and the philosophy of information. The author posits that the current clinical methods for identifying a minimally conscious state in patients based on behavioral assessments may not recognize signs of executive function in post-comatose patients. The author suggests that clinicians should instead look to localized brain "activities in response to task instructions, such as imagining wiggling toes, to diagnose minimal consciousness." The author further suggests that the proposed method is more objective and reliable, because it does not require language comprehension, which may be severely impaired for patients in a minimally conscious state. This paper opens up new perspectives on the philosophy of information as applied philosophy, and as with all good papers, the references provide a detailed review of the related literature.

The discussion around the fundamental issues of the philosophy of information is the topic of a paper by Łukasz Mścisławski (Wrocław University of Science and Technology) entitled "Is Information Ontological or Physical, or Is It Perhaps Something Else? Some Remarks on Krzanowski's Approach to the Concept of Information." The paper presents a critical evaluation of the concept of physical information that has been proposed by Roman Krzanowski. According to Mścisławski, the concept of physical information may play an important role in the philosophy of physics and metaphysics, the philosophy of information, and computer science. The author further

states that the distinctions between ontological, which is another term used to denote physical information, and epistemological "information can be regarded as being analogous to G.F.R. Ellis's analyses of the passage of time in his concept of the Crystallizing Block Universe." For anyone wanting to become familiar with the concept of physical information and its potential implications for cosmology, physics, and computing, this paper is a good place to start.

The next paper in the collection was penned by Kristina Šekrst and Sandro Skansi (University of Zagreb), and it is entitled "Machine learning and essentialism." This paper studies the connection between machine learning and essentialism. The authors posit that similarity-based approaches are more suited for pattern recognition and "complex deep-learning issues, while for classification problems, mostly for unsupervised learning, essentialism seems like the best choice." The authors conclude that essences are not present in data but rather in learned targets, so machine learning does not provide any evidence for the independent existence of essential properties. Thus, our experiences with machine learning, according to the authors, do not offer any proof to support the ontological status of essences. A substantial list of references related to essentialism and machine learning is provided at the end of the paper.

A complementary view about the ontological commitments of artificial intelligence is presented in the paper written by Roman Krzanowski and Paweł Polak (Pontifical University of John Paul II in Kraków) entitled "The Meta-Ontology of AI systems with Human-Level Intelligence." Meta-ontology in philosophy is a discourse centered on ontology, ontological commitment, and the truth condition of ontological theories. The authors posit that the meta-ontology of current AI systems is concerned with computational representations of reality in the form of structures, data constructs, and computa-

tional concepts, while the ontological commitment of AI systems with human-level intelligence must be directed at what exists in the outside world. This paper builds upon the ontological postulates that were formulated by Brian Cantwell Smith about AI systems, and an extensive list of relevant literature is also of course provided.

The final paper was written by Krzysztof Sołoducha (Military University of Technology), and it is titled "Analysis of the Implications of the 'Moral Machine' Project as an Implementation of the Concept of 'Coherent, Extrapolated Volition' for Building Clustered Trust in Autonomous Machines." This paper focuses on performing an "analysis of Eliezer Yudkowsky's concept of 'coherent extrapolated volition' (CEV) as a response to the need for a post-conventional, persuasive morality that meets the criteria of active trust in the sense of Anthony Giddens." In the paper, the "authors formulate guidelines for transformation of the idea of a coherent extrapolated volition into the concept of a coherent, extrapolated and clustered volition."In the author's words, "The argumentation used in the paper is intended to show that the idea of CEV transformed into its clustered version can be used to build a technically and socially efficient decision-making pattern database for autonomous machines." As with any excellent paper, an extensive list of relevant resources is provided.

In addition to the eight abovementioned papers, there are two essays. These differ from the papers by presenting a more open perspective that allows for some personal views that would likely be too tentative for a formal work. The essay format therefore allows authors to share creative ideas beyond formal hypotheses and present the reader with some inspiring and challenging reading.

The first essay by Kazimierz Trzęsicki is titled "Perspective on Turing Paradigm." It argues that Turing planted the seeds of a new paradigm in which the *book of nature* is written in algorithms. In

his arguments, the author delves far into the past, touching upon the works of the Babylonians and Egyptians, as well as later figures like Roger Bacon, Nicolas de Condorcet, Galileo, Leibnitz, and many others. The value of this paper lies in how the author tries to connect all of these past, geographically dispersed thinkers with modern ideas. Nevertheless, the success of this approach should be judged by the reader. The concepts and personalities collected in this essay are so extensive that Turing himself would have been surprised by how many people contributed to his ideas. After all, it is hard to be original!

The second essay was written by Adam Olszewski (Pontifical University of John Paul II in Kraków), and it is titled "Will a Human Always Outsmart a Computer? An Essay." The author presents the model for the "outsmarting" of a machine by a human based on a mathematical game between two players (the base domain), such that winning the game is denoted as "outsmarting." The game in question is similar to a Banach-Mazur game. The author concludes that while in the gaming example, a man beats the hypothetical machine, the question is then this: How far can the results of this thought experiment be generalized? A rather frugal reference list gives sufficient links to sources for those less familiar with the discussed ideas.

Finally, we have three book reviews: The first review by Paweł Polak concerns Roman Krzanowski's (2021) book *Ontological information: information in the physical world* (Hackensack, New Jersey: World Scientific). This review is a sort of addendum to Mścisławski's previously mentioned paper, and it exposes the philosophical underpinnings of Krzanowski's book and the perspectives it opens up. The second review, also by Paweł Polak, is for Andrzej Bielecki's (2019) book *Models of Neurons and Perceptrons: Selected Problems and Challenges* (Cham: Springer International Publishing). Bielecki's work makes important contributions to contemporary philosophy in

science by showing the role of computing in mathematizing subcellular biology. The third book review by Łukasz Mścisławski concerns a book by Cappelen and Dever (2021) entitled *Making AI Intelligible. Philosophical Foundations* (Publishing: Oxford University Press). This book examines possible ways to make AI intelligible, and many questions remain to be asked about this from a philosophical perspective.

With thirteen works in the form of papers, essays, and book reviews, this special edition of ZFN represents a fairly substantial package of ideas and concepts. No one is obliged or expected to read all these works, but whatever essays or papers the reader chooses to digest will likely be greatly rewarding.

Last but not least, we would like to acknowledge the excellent work of this ZFN edition's editors, Paweł Polak and Piotr Urbańczyk. Without their dedication and long nights of effort, this publication would not have been possible.

Some of the papers collected in this edition of ZFN were presented at the Philosophy in Informatics VI: Frontiers of Philosophy of Computing and Information conference held on December 16–17, 2021, and organized by the Polish Academy of Arts and Sciences (PAU). We would like to thank Gordana Dodig-Crnkovic for supporting this conference and contributing to this special edition.

Roman Krzanowski
Editor of this special ZFN collection

Articles

Artykuły

In search of a common, information-processing, agency-based framework for anthropogenic, biogenic, and abiotic cognition and intelligence

Gordana Dodig-Crnkovic
Chalmers University of Technology,
University of Gothenburg
and Mälardalen University, Sweden

Abstract

Learning from contemporary natural, formal, and social sciences, especially from biology, as well as from humanities, particularly contemporary philosophy of nature, requires updates of our old definitions of cognition and intelligence. The result of current insights into basal cognition of single cells and evolution of multicellular cognitive systems within the framework of extended evolutionary synthesis (EES) helps us better to understand mechanisms of cognition and intelligence as they appear in nature. New understanding of information and processes of physical (morphological) computation contribute to novel possibilities that can be used to inspire the development of abiotic cognitive systems (cognitive robotics), cognitive computing and artificial intelligence.

Keywords

information, computation, cognition, intelligence, extended evolutionary synthesis, anthropogenic, biogenic and abiotic cognition.

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 17–46 • CC-BY-NC-ND 4.0

Information, computation, cognition, intelligence, and evolution of living organisms

The notion of information is used nowadays not only to refer to means of communication between humans, but also to denote data structures utilized for communication by other living organisms, even the simplest ones like single cells as used in the fields of bioinformatics or neuroinformatics.

In what follows we build on the ideas presented in (Dodig-Crnkovic, 2017a):

> a view of nature as a network of info-computational agents organized in a dynamical hierarchy of levels. It provides a framework for unification of currently disparate understandings of natural, formal, technical, behavioral, and social phenomena based on information as a structure, differences in one system that cause the differences in another system, and computation as its dynamics, i.e., physical process of morphological change in the informational structure.

In the current definition of computation as a dynamic of information, computation is taken to be any process of information transformation that leads to behavior, and not only those processes that we currently use to calculate, manually or with a machinery:

> Traditionally, the dynamics of computing systems, their unfolding behavior in space and time has been a mere means to the end of computing the function which specifies the algorithmic problem which the system is solving. In much of contemporary computing, the situation is reversed: the purpose of the computing system is to exhibit certain behaviour. [...] We need a theory of the dynamics of informatic processes, of interaction, and information flow, as a basis for answering

> such fundamental questions as: What is computed? What is a process? What are the analogues to Turing completeness and universality when we are concerned with processes and their behaviors, rather than the functions which they compute? (Abramsky, 2008)

Cognition can be defined as a process of "being in the world" of an agent. For living organisms, cognition is a process of life (perception, internal process control by information, actuation/agency) (Maturana, 1970; Maturana and Varela, 1980; Stewart, 1996). Cognition of an organism is based on the ability to learn from the environment and adapt so as to survive as an individual and as a species, for which organisms use information and its processing (computation).

Intelligence, as capacity for problem-solving within an environment/context, can be seen as one of the features of cognition. It is found in all living organisms as they all possess cognition, from single cells to their complex structures constituting tissues, organs, and organisms in constant interaction with each other and with the environment.

Human intelligence is the object of most of the studies of intelligence. Often it is considered to be a multidimensional phenomenon, that includes both classical problem-solving and decision-making ability (logical-mathematical reasoning), existential intelligence (ability to survive), visual-spatial, musical, bodily-kinesthetic, naturalist, linguistic, interpersonal (social), and intra-personal (ability of inner insight) intelligence. However, the question of cognition and intelligence in non-human animals and other organisms is still controversial in philosophy of mind, psychology and even in some parts of cognitive science (Ball, 2022).

Cognition and intelligence on different levels of organization of life—embodied, embedded, enacted, and extended

With the increasing insights into empirical details of processes and structures of cognition, it is emerging that human cognition and intelligence are based not only on activities of nervous system with neurons and glia cells, but equally importantly results from their interaction with non–neuronal subsystems including immune system and other somatic cells as well as the exchanges of the body with the environment. It comes as no surprise, as the nervous system is in a close interaction with the rest of the body.

Human nervous system is made up of two types of cells: primary neurons and glial cells, and it is divided into two parts: the central nervous system (brain and spinal cord) and the peripheral nervous system (autonomic and somatic nervous systems). The nervous system controls and regulates the activities of organs and systems through neuron feedback, enabling the body to respond to environmental changes (Biotechnology-Accegen, 2022). Through the embodiment, the nervous system also communicates with the external world, including other cognitive agents. The understanding that human cognition results from the activities of different types of cells, and not only nerve cells (neurons), is based on the new recognition of the existence of basal cognition/ minimal cognition / microorganismic cognition and intelligence. Unicellular organisms (single cells) have sensors and actuators and use chemical signaling and transfer of genetic information as a basis for adaptation and learning (Baluška and Levin, 2016; Ng and Bassler, 2009; Witzany, 2011; Ben-Jacob, 2003; Ben-Jacob, Shapira and Tauber, 2006). Cognitive (sensory-based) and intelligent

(problem-solving) processes are regulating the state of a single cell which is a building block of multicellular living organisms (Manicka and Levin, 2019).

Thus, recently the ideas of cognition and intelligence have increased in scope (Dennett, 2017) with improved understanding of their underlying mechanisms—from the activity on the level of the human brain, to the processes on the somatic cell level. Single cell need not be a part of a human body to be seen as performing cognitive and intelligent behavior, it could be a unicellular organism or a constituent part of an animal or a plant.

At the same time as new insights have been made into the nature of biological cognition, computational and robotic cognitive systems are being developed with various degrees of cognition and intelligence. Some functions of artificial intelligence surpass human capacities (such as processing parallelism, search, memory, precision, and correctness, and often also speed) while many other capacities are far below the human level, such as common-sense reasoning, or goal-directed agency in the sense of self-preservation and self-organization.

Understanding cognition and intelligence in nature on different levels of organization, because of their fundamentally biological mechanisms is only possible if we see it in the context of evolution. As in all of biology, "nothing makes sense except for in the light of evolution" (Dobzhansky, 1973), and the cognition as a process can only be understood in the light of evolution.

However, new *abiotic approaches to cognition* assume that it is possible to construct cognitive agents from abiotic elements. Artificial (artifactual) intelligence is an attempt to produce intelligent behaviors akin to those shown by living beings (from the beginning specifically

in humans) but implemented in non-living substrate. We can compare "cognitive behavior" of abiotic systems with the cognitive behavior of living organisms and see how close they are to each other.

The necessity of the Extended evolutionary synthesis (EES)

Looking at the intelligence of a living organism as a result of information processing and embodied goal-directed behaviors on hierarchy of levels of organization, suggests necessity of understanding of the process of evolution in a broader and more inclusive way than before, where biological agents are seen in their natural environments, from single cells to groups of organisms. A scientific meeting organized in partnership with the British Academy by Denis Noble, Nancy Cartwright, Patrick Bateson, John Dupré and Kevin Laland presented and discussed those important *New trends in evolutionary biology in biological, philosophical, and social science perspectives* (Royal Society, 2016).

That emerging view of evolution is called Extended Evolutionary Synthesis (EES), which is a new interpretation of the theory of evolution based on the latest scientific knowledge about life and its changes, emphasizing fundamental mechanisms of constructive development and reciprocally causal nature between an organism and its environment (Schwab, Casasa and Moczek, 2019) More on Extended Evolutionary Synthesis can be found in (Laland et al., 2015), presenting EES and its structure, assumptions, and predictions, and (Müller, 2017a,b) explaining why an extended evolutionary synthesis is necessary. Svensson (2018) argues:

> The Extended Evolutionary Synthesis (EES) will supposedly expand the scope of the Modern Synthesis (MS) and Standard Evolutionary Theory (SET), which has been characterized as gene-centered, relying primarily on natural selection and largely neglecting reciprocal causation.

Evolution is a result of interactions between natural agents, cells and their groups on variety of levels of organization (Jablonka and Lamb, 2014; Laland et al., 2015; Ginsburg and Jablonka, 2019), as Jablonka and Lamb argue in their book *Evolution in Four Dimensions*: *Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. These dimensions can be found on different levels of organization of life.

In short, if we want to bring evolutionary theory in coherence with the advancement in other sciences, extended evolutionary synthesis is necessary.

Info-computational lens: agent-based natural information and computation

We use an info-computational lens to approach phenomena of cognition and intelligence. A framework of (Dodig-Crnkovic, 2017c) enables understanding of cognitive systems generated through self-structuring processes of morphological info-computation on the hierarchy of levels in nature from physics, to chemistry and biology, based on agent-centric embodied information and morphological computation. It means that we assume:

- computing nature paradigm, where nature is seen through the lens of information and computation as its dynamics, that is

providing a basis for unification of currently disparate understanding of natural, formal, technical, social and behavioral phenomena;
- an observer-dependent, agent-based reality, that is reality for an agent for which cognition is a result of relational info-computational processes;
- computational interpretation of information dynamics in nature, where computation is physical (morphological) computation;

that enables us to:

- avoid frequent misunderstandings of the inadequate abstract models of computation (as in old computationalism) and focus on embodied morphological computation in physical systems, especially cognitive ones such as living beings;
- suggest the necessity of generalization of the models of computation beyond the traditional Turing machine model and acceptance of "second generation" models of computation capable of covering the whole range of phenomena from physics to cognition (Abramsky);
- understand goal directed behaviors and complexification in living systems through the extended evolutionary synthesis.

The developments supporting info-computational approach, as a variety of naturalism, are found in among others complexity theory, systems theory, theory of computation (natural computing, organic computing, unconventional computing), cognitive science, neuroscience, information physics, agent based models of social systems and information sciences, robotics (especially developmental robotics), bioinformatics and artificial life (Dodig-Crnkovic and Müller, 2011;

Dodig-Crnkovic, 2017c), as well as biosemiotics (Sarosiek, 2021) and Polak-Krzanowski's deanthropomorphized pancomputationalism (Polak and Krzanowski, 2019).

Cognition of a living organism is thus studied as a network of networks of distributed information processing units on variety of levels of organization, from single cells to the whole body including the level of groups of organisms manifest as social cognition.

Natural cognition based on cells processing (computing) information—basal cognition in an extended evolutionary perspective

Despite decades of research into the subject, there is still no agreement about where cognition is found in the living world (Ball, 2022). Is a nervous system needed? If so, why? If not, why not? A new two-part theme issue of *Phil Trans B* on the emerging field of 'Basal Cognition', edited by Pamela Lyon, Fred Keijzer, Detlev Arendt and Michael Levin, explores these questions (Levin et al., 2021; Lyon et al., 2021).

Present increase of knowledge about cellular cognition and new gained details of complex goal-directed behaviors is nicely illustrated by the example of a single-celled predator organism *Lacrymaria olor* ("tears of a swan") hunting down another cell, often used by Michael Levin. *Lacrymaria* has a "neck" a "body" and a "mouth". It beats the hair-like cilia around its "head" and extends its neck up to 8 times its body length, while chasing and finally swallowing another cell. It has no nervous system and no sensors that macroscopic living organisms typically use to chase their prey. How does it manage to identify, follow, and catch the prey? The mechanisms

that enable *Lacrymaria* to hunt down another cell, that goal-directedly activate cilia, take care of timing of "mouth" opening and closing are studied in (Weiss, 2020). Likewise, (Coyle et al., 2019) describe how coupled active systems encode an emergent hunting behavior in *Lacrymaria olor*. Even the work of (Mearns et al., 2020) analyzes its hunting behavior, revealing a tightly coupled stimulus-response loop. Furthermore (Wlotzka and McCaskill, 1997) argue that in this case, they observed behavior of a molecular predator and its prey, through coupled isothermal amplification of nucleic acids. In short, research shows that a goal-directed behavior of *Lacrymaria olor*, is a result of a coupled stimulus-response loops. However, importantly, we do not know the meta-level mechanism which activates those loops and makes them goal directed.

Another microorganism under intense study for their goal-directed, efficient learning and adaptive behavior, which are of special interest because of their ability to cause diseases in other organisms, are bacteria. Eshel Ben Jacob have been studying bacterial colonies, their self-organization, complexification and adaptation, smartness, communication and linguistic communication (by chemical languages), social intelligence, natural information processing, and foundations of bacterial cognition (Ben-Jacob, 1998; 2003; 2008; 2009; Ben-Jacob, Becker and Shapira, 2004; Ben-Jacob, Shapira and Tauber, 2006; 2011). Works (Witzany, 2011; Schauder and Bassler, 2001; Waters and Bassler, 2005; Ng and Bassler, 2009) focus on communication (information exchange) mechanisms in bacteria, and especially *quorum sensing*, where group of bacteria make a majority-based decisions. Bacteria colonies and films display various multicellular behaviors, emitting, receiving, and processing a large vocabulary of chemical symbols.

More about experimental methods for study of cell cognition can be found in the work of *The cell cognition project* (Held et al., 2010).

From all the above evidence it is clear that unicellular organisms exhibit basal cognition and intelligence (problem-solving capacities). A fundamental observation connecting this rudimentary biotic cognition and more complex anthropogenic (i.e., human-level, brain-based) cognition, is the following:

> 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 (Lyon et al., 2021).

Similar arguments for biogenic nature of cognition have been presented by (Levin et al., 2021; Yuste and Levin, 2021; Lyon et al., 2021).

Our approach to information-processing mechanisms of cognition, unlike vast majority of artificial cognitive architectures targeting human-level cognition, focus on the development and evolution of the continuum of natural cognitive architectures, from basal cellular architecture up, as studied by (Levin et al., 2021) and already identified by (Sloman, 1984).

The connection between high-level and basal cognition is visible in the role of ion channels and neurotransmitters, studied in nervous cells, but also present in ordinary somatic cells:

> We have previously argued that the deep evolutionary conservation of ion channel and neurotransmitter mechanisms highlights a fundamental isomorphism between developmental and behavioral processes. Consistent with this, membrane excitability has been suggested to be the ancestral basis for

> psychology [...]. Thus, it is likely that the cognitive capacities of advanced brains lie on a continuum with, and evolve from, much simpler computational processes that are widely conserved at both the functional and mechanism (molecular) levels.
>
> The information processing and spatio–temporal integration needed to construct and regenerate complex bodies arises from the capabilities of single cells, which evolution exapted and scaled up as behavioral repertoires of complex nervous systems that underlie familiar examples of Selves (Fields and Levin, 2019).

This biogenic nature of cognition makes it necessary to recognize all living forms, and not only those with nervous systems (Piccinini, 2020), or what is even more frequent only humans, as cognitive systems.

As for the driving mechanisms behind this complexification process in living/cognitive systems, (Fields, Friston et al., 2022, pp.1–2) describe how The Free Energy Principle of Karl Friston can drive neuromorphic development in the fully-general quantum-computational framework of topological quantum neural networks:

> We show how any system with morphological degrees of freedom and locally limited free energy will, under the constraints of the free energy principle, evolve toward a neuromorphic morphology that supports hierarchical computations in which each "level" of the hierarchy enacts a coarse-graining of its inputs, and dually a fine-graining of its outputs. Such hierarchies occur throughout biology, from the architectures of intracellular signal transduction pathways to the large-scale organization of perception and action cycles in the mammalian brain.

Biogenic approach is useful not only for understanding of cognition and intelligence and their evolution in living nature, but also for engineering of artificial systems that need certain level of intelligence, not necessarily the human level, such as nano-bots (Kriegman et al., 2021) or different elements of IoT (Internet of Things).

Cognitive computing and AI—still anthropogenic

Inspired by the behaviors produced by anthropogenic cognition, (Modha et al., 2011) the field of cognitive computing is exploring biomimetic approaches to cognition in abiotic systems (Gudivada et al., 2019) studying cognitive computing systems, their potential and possible futures. In the application domain, e.g. IBM had a cognitive computing project called Systems of Neuromorphic Adaptive Plastic Scalable Electronics (SyNAPSE) (Srinivasa and Cruz-Albrecht, 2012).

The quest for intelligent machines ultimately requires new breakthroughs in computer architecture, theory of computation, computational neuroscience, supercomputing, cognitive science, and related fields orchestrated in a coherent, unified effort.

Cognitive computing, AI and cognitive robotics present attempts to construct abiotic systems exhibiting cognitive characteristics of biotic systems. As a rule, they assume human-level intelligence and human-level cognition, even though biogenic approaches would bring huge benefits. When we acknowledge that cognition in living nature comes in degrees, it is more meaningful to talk about cognition of artifacts, even though the role of cognitive capacities for an artefact is not to assure its continuing existence (unlike in cognition = life

(Stewart, 1996), which gives the evolutionary role to cognition in biotic systems). The difference is that cognitive artifacts are constructed to pursue human goals, not their own intrinsic ones.

Cognition at different levels of organization of a living organism—from cells up

Traditional anthropogenic approach to cognition (Markram, 2012) is looking at cognition and intelligence in humans as the only natural cognitive agents.

Biogenic approaches on the other hand broaden the domain, seeing cognition as an ability of all living organisms (Maturana, 1970; Maturana and Varela, 1980; Stewart, 1996).

More specifically, Maturana and Varela argue:

> A cognitive system is a system whose organization defines a domain of interactions in which it can act with relevance to the maintenance of itself, and the process of cognition is the actual (inductive) acting or behaving in this domain. *Living systems are cognitive systems and living as a process is a process of cognition.* This statement is valid for all organisms, with and without a nervous system (Maturana and Varela, 1980, p.13; cf. Maturana and Varela, 1992).

Cognition is thus a capacity possessed in different forms and degrees of complexity by every living organism. It is entirety of processes going on in an organism that keeps it alive, and present as a distinct agent in the world. A single cell while alive constantly cognizes, that is registers inputs from the world and its own body, ensures its own continuous existence through metabolism and food hunting while avoiding dangers that could cause its disintegration or damage, at

the same time adapting its own morphology to the environmental constraints. The entirety of physico-chemical processes depends on the morphology of the organism, where morphology is meant as the form and structure. Work of Marijuán, Navarro and del Moral (2010) presents a study of prokaryotic intelligence and its strategies for sensing the environment.

Multicellularity

Unicellular organisms such as bacteria communicate and build swarms or films with far more advanced capabilities compared to individual organisms, through social (distributed) cognition.

In general, groups of smaller organisms (cells) in nature cluster into bigger ones (multicellular assemblies) with differentiated control mechanisms from the cell level to the tissue, organ, organism and groups of organisms, and this layered organization provides information processing benefits.

Examining the origin of multicellularity (Fields and Levin, 2019) investigates the computational boundary of a "self" and argues that it is bioelectricity that drives multicellularity and scale-free cognition. According to (Fields and Levin, 2019), somatic multicellularity presents a satisficing solution to the prediction-error minimization problem for single cells. From the point of view of information, (Colizzi, Vroomans and Merks, 2020) argue that evolution of multicellularity results from a collective integration of spatial information, while (McMillen, Walker and Levin, 2022) show how to use Shannon information theory (Shannon, 1948) as a tool for integration of biophysical signaling modules. By mapping information flow between cells and

pathways, researchers show that information theory supports systems-level view of biological phenomena where molecular reductionism does not work well.

Computationalism is not what it used to be …

… that is the thesis that human cognition and intelligence are Turing machines (Scheutz, 2002). Unlike classical computationalism based on symbol manipulation and Turing Machine model, modern computationalism for modelling of cognitive processes requires new models of computation.

Turing Machine is an abstract logical model of computation equivalent to an algorithm, and it may be used for description of elementary sequential processes in living organisms. However, complex networked physical processes with temporal and other physical resource constraints cannot be adequately modelled as series of sequential logical operations (Turing machines). As Leslie Valiant (2013) succinctly puts it:

> We need computational models for the basic characteristics of life, such as the ability to differentiate and synthesize information, make a choice, adapt, evolve, and learn in an unpredictable world. That requires computational mechanisms and models which are not "certainly, exactly correct" and predefined as Turing machine, but, instead, "probably approximately correct" (PAC).

Computational approaches that are capable of modelling adaptation, evolution and learning are found in the field of natural computation and computing nature (Dodig-Crnkovic, 2014a).

Computing, the fourth scientific domain

Info-computational approach incorporates our best current scientific knowledge about the processes in nature, translating them into language of natural information and computation.

The aim of this approach to cognition is to increase understanding of cognitive capacities in diverse types of agents, biological and synthetic, including their ability of learning, and learning to learn (meta-learning) (Dodig-Crnkovic, 2020) as well as their communication and mutual interactions. According to (Denning, 2007), computing can be seen as a natural science. Even more than that, we are witnessing the emergence of a new computing science (Denning, 2010) , connecting natural and formal sciences, adding the dimension of real time and physical constraints to logic and mathematics. As Rosenbloom argues, "Computing may be the fourth great domain of science along with the physical, life and social sciences" (Rosenbloom, 2015). In that new broader, emerging computing science, the Turing Model of computation is a proper subset.

Computing nature and nature inspired computation. Self-generating systems

Complex biological systems must be modeled as self-referential, self-organizing "component-systems" (Kampis, 1991) which are self-generating and whose behavior, though computational in a general sense, goes far beyond Turing machine model. Georg Kampis studied the behavior of self-modifying systems in biology and cognitive science as a basis for a new framework for dynamics, information, computation, and complexity:

> a component system is a computer which, when executing its operations (software) builds a new hardware. [... We] have a computer that re-wires itself in a hardware-software interplay: the hardware defines the software, and the software defines new hardware. Then the circle starts again (Kampis, 1991, p.223).

Similar position is presented in (Dodig-Crnkovic and Müller, 2011) connecting models of computation from the formal sequential logical machine Turing model to the physical (morphological) concurrent natural computation.

Evolution as generative mechanism for increasingly complex cognitive systems

New insights about cognition and its evolution and development in nature, from cellular to human cognition can be modelled as natural information processing/ natural computation/ morphological computation. In the info-computational approach, evolution in the sense of Extended evolutionary synthesis is a result of interactions between natural agents, cells, and their groups.

Evolution provides generative mechanisms for the emergence of more and more competent living organisms, with increasingly complex natural cognition and intelligence, and those mechanisms can be used as a template for the design and construction of their artifactual, computational counterparts.

Learning from biogenic computing

The concept of biological computation posits that living organisms process information and thus perform computations, and that ideas of information and computation are the key to understanding, modeling, simulation, and control of biological systems. See (Mitchell, 2012) for the exposition of the concept of biological computation, and (Dodig-Crnkovic, 2022) for presentation of cognitive architectures based on natural infocomputation. Cognition as a result of information processing in living agent's morphology, with species-specific cognition and intelligence is described in (Dodig-Crnkovic, 2021).

One of important characteristics of natural computing is its computational efficiency which is becoming increasingly important in the world with pervasive computing and concurrent global warming. The Turing Machine model of computation is not resource-aware, unlike living systems which are constantly optimizing their use of natural resources. Therefore, in the biomimetic approach to cognitive architectures designers are learning from nature how to compute more resource efficiently. Mutual learning between computing, cognitive sciences and neurosciences (Rozenberg and Kari, 2008) leads to improved understanding of how cognition works and develops in nature, and how we can simulate, emulate, and engineer abiotic cognition and intelligence with the properties close to the biotic one.

Morphological computation connecting body, brain, and environment in robotics

The research performed in the diverse fields of soft robotics / self-assembly systems and molecular robotics / self-assembly systems

at all scales / embodied robotics / reservoir computing / real neural systems / systems medicine / functional architecture / organization / process management / computation based on spatio-temporal dynamics/ information theoretical approach to embodiment mechatronics / amorphous computing / molecular computing – all connect body, control ("brain") and environment.

In robotics, a brain and body that researchers learn from, sometimes belongs to an octopus, which unlike typical robots has soft body that presents substantially different possibilities from rigid bodies of conventional robots.

Pfeifer and Bongard (2006) were among the first to present a new view of embodied intelligence, arguing that the body shapes the way we think, looking in the first place from the anthropocentric perspective, but the approach applies equally well to biocentric view of cognition. In biologically inspired robotics, embodiment and self-organization are driving forces of evolving intelligence (Pfeifer, Lungarella and Iida, 2007). They are best understood in terms of morphological computation (Pfeifer and Iida, 2005; Hauser, Füchslin and Pfeifer, 2014).

The essential property of morphological computation is that it is defined on a structure of nodes (agents) that exchange (communicate) information. It is thus applied not only in robotics, but generalized to other physical information-processing systems, including living beings (Dodig-Crnkovic, 2013b; 2017b; 2018).

Computing Nature and Natural Computation

In his article "Epistemology as Information Theory", Greg Chaitin argues that knowledge should be studied as a result of information processes, thus turning epistemology into study of information:

> And how about the entire universe, can it be considered to be a computer? Yes, it certainly can, it is constantly computing its future state from its current state, it's constantly computing its own time-evolution! And as I believe Tom Toffoli pointed out, actual computers like your PC just hitch a ride on this universal computation! (Chaitin, 2007, p.13)

David Deutsch in his article "What is Computation? (How) Does Nature Compute?" contributes with the similar position in the book "A Computable Universe" by Hector Zenil (2012).

Starting from the above ideas, (Dodig-Crnkovic, 2007) proposes that epistemology can be naturalized through the info-computational approach to knowledge generation. The computing nature framework (naturalist computationalism) makes it possible to describe all cognizing agents (living organisms and artificial cognitive systems) as informational structures with computational dynamics (Dodig-Crnkovic and Burgin, 2011; Dodig-Crnkovic, 2013a; 2014b; 2017a; Dodig-Crnkovic and Giovagnoli, 2013; 2017). Morphological computation in this framework is a process of creation of new informational structures, as it appears in nature, living as non-living. It is a process of morphogenesis, which in biological systems is driven by development and evolution (Dodig-Crnkovic, 2013b; 2017b; 2018).

It is worth noting that research on "computing nature" focuses on how physical/ natural/ morphological processes can be interpreted as computation and used to compute, while research on "computable universe" asks the question if we can compute (with our current theories of computing) what we observe as the universe—two different research programs.

Conclusions

New insights from complexity theory, systems theory, theory of computation (natural computing, organic computing, unconventional computing), cognitive science, neuroscience, information physics, agent based models of social systems and information sciences, robotics, as well as bioinformatics and artificial life call for updates in our understanding of cognition and intelligence (Dodig-Crnkovic and Müller, 2011; Dodig-Crnkovic, 2017c).

Traditionally, in the fields of cognitive science, philosophy of mind, cognitive computing and artificial intelligence, cognition and intelligence are assumed to be the abilities of humans. They are described in terms of concepts such as mind, thought, reasoning, logic, etc. However, new understanding of the goal-directed, learning, and adaptive behaviors of all living organisms, from all five kingdoms of life—animal, plant, fungi, protist and monera, from single celled to multi-cellular organisms and their ecologies, all possess level of cognition and intelligence which increases with the complexity of the system.

In this article we present a common framework of info-computation, where computation is physical/morphological computation providing unified approach to anthropogenic, biogenic, and abiotic cognition. The advantage of info-computational approach is that it enables learning of mechanisms of those three types of cognition and intelligence. It also connects different levels of organization as observed in nature.

Cognition and intelligence, coming from the simplest to the most complex in a continuum of natural systems can be source of inspiration

for the design and construction of artificial cognitive systems with varying degrees/levels of intelligence, from nano-bots to autonomous cars and android robots.

Bibliography

Abramsky, S., 2008. Information, Processes and Games. In: J. Benthem van and P. Adriaans, eds. *Philosophy of Information*. Amsterdam, The Netherlands: North Holland, pp.483–549.

Ball, P., 2022. *The Book of Minds: How to Understand Ourselves and Other Beings, from Animals to AI to Aliens*. Chicago: University of Chicago Press.

Baluška, F. and Levin, M., 2016. On having no head: cognition throughout biological systems. *Frontiers in Psychology*, 7, p.902.

Ben-Jacob, E., 1998. Bacterial wisdom, Gödels theorem and creative genomic webs. *Physica A*, 248, pp.57–76.

Ben-Jacob, E., 2003. Bacterial Self-Organization: Co-Enhancement of Complexification and Adaptability in a Dynamic Environment. *Phil. Trans. R. Soc. Lond. A*, pp.315–322.

Ben-Jacob, E., 2008. Social behavior of bacteria: from physics to complex organization. *The European Physical Journal B*, 65(3), pp.315–322.

Ben-Jacob, E., 2009. Learning from Bacteria about Natural Information Processing. *Annals of the New York Academy of Sciences*, 1178, pp.78–90.

Ben-Jacob, E., Becker, I. and Shapira, Y., 2004. Bacteria Linguistic Communication and Social Intelligence. *Trends in Microbiology*, 12(8), pp.366–372.

Ben-Jacob, E., Shapira, Y. and Tauber, A., 2006. Seeking the Foundations of Cognition in Bacteria. *Physica A*, 359, pp.495–524.

Ben-Jacob, E., Shapira, Y. and Tauber, A., 2011. Smart Bacteria. In: L. Margulis, C. Asikainen and W. Krumbein, eds. *Chimera and Consciousness. Evolution of the Sensory Self*. Cambridge; Boston: MIT Press.

Biotechnology-Accegen, 2022. *Nervous System Primary Cells* [Online]. Available at: <https://www.accegen.com/category/nervous-system-primary-cells/> [visited on 10 January 2023].

Chaitin, G., 2007. Epistemology as Information Theory: From Leibniz to ω. In: G. Dodig Crnkovic, ed. *Computation, Information, Cognition – The Nexus and The Liminal*. Newcastle UK: Cambridge Scholars Pub., pp.2–17.

Colizzi, E.S., Vroomans, R.M. and Merks, R.M., 2020. Evolution of multicellularity by collective integration of spatial information. *eLife* [Online], 9:e56349. https://doi.org/10.7554/eLife.56349.

Coyle, S.M. et al., 2019. Coupled Active Systems Encode an Emergent Hunting Behavior in the Unicellular Predator Lacrymaria olor. *Current Biology* [Online], 29, pp.3838–3850. https://doi.org/10.1016/j.cub.2019.09.034.

Dennett, D., 2017. *From Bacteria to Bach and Back: The Evolution of Minds*. New York: Norton & Company.

Denning, P., 2007. Computing is a natural science. *Communications of the ACM*, 50(7), pp.13–18.

Denning, P., 2010. Computing Science: The Great Principles of Computing. *American Scientist*, 98(5), pp.369–372.

Dobzhansky, T., 1973. Nothing in Biology Makes Sense Except in the Light of Evolution. *American Biology Teacher* [Online], 35(3), pp.125–129. https://doi.org/10.2307/4444260.

Dodig-Crnkovic, G., 2007. Epistemology Naturalized: The Info-Computationalist Approach. *APA Newsletter on Philosophy and Computers*, 06(2), pp.9–13.

Dodig-Crnkovic, G., 2013a. The Development of Models of Computation with Advances in Technology and Natural Sciences. *6th AISB Symposium on Computing and Philosophy: The Scandal of Computation - What is Computation? - AISB Convention 2013*.

Dodig-Crnkovic, G., 2013b. The Info-computational Nature of Morphological Computing. In: V.C. Müller, ed. *Philosophy and Theory of Artificial*

Intelligence [Online]. Vol. 5, *Studies in Applied Philosophy, Epistemology and Rational Ethics*. Berlin, Heidelberg: Springer, pp.59–68. https://doi.org/10.1007/978-3-642-31674-6_5.

Dodig-Crnkovic, G., 2014a. Modeling Life as Cognitive Info-Computation. In: A. Beckmann, E. Csuhaj-Varjú and K. Meer, eds. *Computability in Europe 2014. LNCS*. Berlin; Heidelberg: Springer, pp.153–162.

Dodig-Crnkovic, G., 2014b. Why we need info-computational constructivism. *Constructivist Foundations*, 9(2), pp.246–255.

Dodig-Crnkovic, G., 2017a. Computational Dynamics of Natural Information Morphology, Discretely Continuous. *Philosophies* [Online], 2(4), p.23. https://doi.org/10.3390/philosophies2040023.

Dodig-Crnkovic, G., 2017b. *Morphologically Computing Embodied, Embedded, Enactive, Extended Cognition* [Online]. PT-AI 2017. Available at: <https://www.pt-ai.org/2017/posters>.

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

Dodig-Crnkovic, G., 2018. Cognition as Embodied Morphological Computation. In: V.C. Müller, ed. *Philosophy and Theory of Artificial Intelligence 2017* [Online], Studies in Applied Philosophy, Epistemology and Rational Ethics. Cham: Springer International Publishing, pp.19–23. https://doi.org/10.1007/978-3-319-96448-5_2.

Dodig-Crnkovic, G., 2020. Natural Morphological Computation as Foundation of Learning to Learn in Humans, Other Living Organisms, and Intelligent Machines. *Philosophies* [Online], 5(3), pp.17–32. https://doi.org/10.3390/philosophies5030017.

Dodig-Crnkovic, G., 2021. Cognition as a Result of Information Processing in Living Agent's Morphology. Species-specific Cognition and Intelligence. *Proceedings of 16th SweCog Conference* [Online], pp.22–25. Available at: <https://www.diva-portal.org/smash/get/diva2:1611781/FULLTEXT01.pdf> [visited on 10 January 2023].

Dodig-Crnkovic, G., 2022. Cognitive architectures based on natural info-computation. In: V.C. Müller, ed. *Philosophy and theory of artificial intelligence 2021*. SAPERE. Berlin, Heidelberg: Springer.

Dodig-Crnkovic, G. and Burgin, M., 2011. *Information and Computation*. Singapore: World Scientific.

Dodig-Crnkovic, G. and Giovagnoli, R., 2013. Computing nature – A network of networks of concurrent information processes. *Computing Nature* [Online]. Vol. 7. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-37225-4_1.

Dodig-Crnkovic, G. and Giovagnoli, R., 2017. *Representation and Reality in Humans, Other Living Organisms and Intelligent Machines* [Online]. Ed. by G. Dodig-Crnkovic and R. Giovagnoli, *Studies in Applied Philosophy, Epistemology and Rational Ethics, 28*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-43784-2.

Dodig-Crnkovic, G. and Müller, V.C., 2011. A Dialogue Concerning Two World Systems: Info-Computational vs. Mechanistic. In: G. Dodig Crnkovic and M. Burgin, eds. *Information and Computation*. Singapore: World Scientific Pub Co Inc, pp.149–184.

Fields, C., Friston, K. et al., 2022. *The Free Energy Principle drives neuromorphic development* [Online]. arXiv. Available at: <http://arxiv.org/abs/2207.09734> [visited on 10 January 2023].

Fields, C. and Levin, M., 2019. Somatic multicellularity as a satisficing solution to the prediction-error minimization problem. *Communicative & Integrative Biology* [Online], 12(1), pp.119–132. https://doi.org/10.1080/19420889.2019.1643666.

Ginsburg, S. and Jablonka, E., 2019. *The Evolution of the Sensitive Soul* [Online]. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/11006.001.0001.

Gudivada, V., Pankanti, S., Seetharaman, G. and Zhang, Y., 2019. "Cognitive Computing Systems: Their Potential and the Future". *Computer* [Online], 52(05), pp.13–18. https://doi.org/10.1109/MC.2019.2904940.

Hauser, H., Füchslin, R. and Pfeifer, R., 2014. *Opinions and Outlooks on Morphological Computation*. e-book.

Held, M. et al., 2010. CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging. *Nature Methods*, 7(9), pp.747–754.

Jablonka, E. and Lamb, M., 2014. *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life. Revised Edition. Life and Mind: Philosophical Issues in Biology and Psychology*. Cambridge, MA: A Bradford Book; MIT Press.

Kampis, G., 1991. *Self-Modifying Systems in Biology and Cognitive Science: A New Framework for Dynamics, Information, and Complexity*. Amsterdam: Pergamon Press.

Kriegman, S., Blackiston, D., Levin, M. and Bongard, J., 2021. Kinematic self-replication in reconfigurable organisms. *Proceedings of the National Academy of Sciences* [Online], 118(49), e2112672118. https://doi.org/10.1073/pnas.2112672118.

Laland, K.N. et al., 2015. The extended evolutionary synthesis: its structure, assumptions and predictions. *Proceedings of the Royal Society B: Biological Sciences* [Online], 282(1813), p.20151019. https://doi.org/10.1098/rspb.2015.1019.

Levin, M., Keijzer, F., Lyon, P. and Arendt, D., 2021. Basal cognition: multicellularity, neurons and the cognitive lens, Special issue, Part 2. *Phil. Trans. R. Soc. B*, 376(20200458).

Lyon, P., Keijzer, F., Arendt, D. and Levin, M., 2021. Basal cognition: conceptual tools and the view from the single cell - Special issue, Part 1. *Phil. Trans. R. Soc. B*, 376(20190750).

Manicka, S. and 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).

Marijuán, P.C., Navarro, J. and del Moral, R., 2010. On prokaryotic intelligence: Strategies for sensing the environment. *BioSystems* [Online], 99(2), pp.94–103. https://doi.org/10.1016/j.biosystems.2009.09.004.

Markram, H., 2012. The Human Brain Project. *Scientific American*, 306(6), pp.50–55.

Maturana, H., 1970. *Biology of Cognition*. (Biological Computer Laboratory Research Report BCL 9). Urbana, IL: University of Illinois.

Maturana, H. and Varela, F., 1980. *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht: D. Reidel Pub. Co.

Maturana, H.R. and Varela, F.J., 1992. *The Tree of Knowledge: The Biological Roots of Human Understanding*. Rev. Boston, MA: Shambala.

McMillen, P., Walker, S.I. and Levin, M., 2022. Information Theory as an Experimental Tool for Integrating Disparate Biophysical Signaling Modules. *Int. J. Mol. Sci.* [Online], 23(9580). https://doi.org/10.3390/ijms23179580.

Mearns, D.S. et al., 2020. Deconstructing Hunting Behavior Reveals a Tightly Coupled Stimulus-Response Loop. *Current Biology* [Online], 30(1), 54–69.e1–e9. https://doi.org/10.1016/j.cub.2019.11.022.

Mitchell, M., 2012. Biological computation. *Computer Journal*, 55(7), pp.852–855.

Modha, D.S. et al., 2011. Cognitive Computing. *Communications of the ACM*, 54(8), pp.62–71.

Müller, G.B., 2017a. Correction to 'Why an extended evolutionary synthesis is necessary'. *Interface Focus*, 7(20170065).

Müller, G.B., 2017b. Why an extended evolutionary synthesis is necessary. *Interface Focus*, 7(2017001520170015).

Ng, W.-L. and Bassler, B., 2009. Bacterial quorum-sensing network architectures. *Annual Review of Genetics*, 43, pp.197–222.

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 Iida, F., 2005. Morphological computation: Connecting body, brain and environment. *Japanese Scientific Monthly*, 58(2), pp.48–54.

Pfeifer, R., Lungarella, M. and Iida, F., 2007. Self-organization, embodiment, and biologically inspired robotics. *Science*, 318, pp.1088–1093.

Piccinini, G., 2020. *Neurocognitive Mechanisms: Explaining Biological Cognition*. Oxford: Oxford scholarship online.

Polak, P. and Krzanowski, R., 2019. Deanthropomorphized Pancomputationalism and the Concept of Computing. *Foundations of Computing and Decision Sciences* [Online], 44(1), pp.45–54. https://doi.org/10.2478/fcds-2019-0004.

Rosenbloom, P., 2015. *On Computing: The Fourth Great Scientific Domain*. Cambridge, MA: MIT Press.

Royal Society, 2016. *New trends in evolutionary biology: biological, philosophical and social science perspectives* [Online]. Available at: <https://royalsociety.org/science-events-and-lectures/2016/11/evolutionary-biology/>.

Rozenberg, G. and Kari, L., 2008. The many facets of natural computing. *Communications of the ACM* [Online], 51, pp.72–83. https://doi.org/10.1145/1400181.1400200.

Sarosiek, A., 2021. The role of biosemiosis and semiotic scaffolding in the processes of developing intelligent behaviour. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* [Online], (70), pp.9–44. Available at: <https://zfn.edu.pl/index.php/zfn/article/view/535>.

Schauder, S. and Bassler, B., 2001. The languages of bacteria. *Genes & Dev.*, 15, pp.1468–1480.

Scheutz, M., 2002. *Computationalism New Directions*. Cambridge, MA: MIT Press.

Schwab, D.B., Casasa, S. and Moczek, A., 2019. On the Reciprocally Causal and Constructive Nature of Developmental Plasticity and Robustness. *Frontiers in Genetics* [Online], 9(735). https://doi.org/10.3389/fgene.2018.00735.

Shannon, C.E., 1948. A mathematical theory of communication. *The Bell System Technical Journal* [Online], 27(July 1928), pp.379–423. https://doi.org/10.1145/584091.584093.

Sloman, A., 1984. The structure of the space of possible minds. In: S. Torrance, ed. *The Mind and the Machine: philosophical aspects of Artificial Intelligence*. Chichester: Ellis Horwood, pp.35–42.

Srinivasa, N. and Cruz-Albrecht, J., 2012. Neuromorphic adaptive plastic scalable electronics: analog learning systems. *IEEE Pulse* [Online], 3(1), pp.51–56. https://doi.org/10.1109/MPUL.2011.2175639.

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

Svensson, E.I., 2018. On Reciprocal Causation in the Evolutionary Process. *Evol Biol*, 45, pp.1–14.

Valiant, L., 2013. *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. New York, NY: Basic Books.

Waters, C.M. and Bassler, B., 2005. Quorum Sensing: Cell-to-Cell Communication in Bacteria. *Annual Review of Cell and Developmental Biology*, 21, pp.319–346.

Weiss, J., 2020. *Single-celled Lacrymaria olor Hunts Down Another Cell* [Online]. Available at: <https://www.youtube.com/watch?v=sq6Y54mxjOg>.

Witzany, G., 2011. Introduction: Key Levels of Biocommunication of Bacteria. In: G. Witzany, ed. *Biocommunication in Soil Microorganisms* [Online]. Vol. 23. Berlin, Heidelberg: Springer, pp.1–34. https://doi.org/10.1007/978-3-642-14512-4_1.

Wlotzka, B. and McCaskill, J.S., 1997. A molecular predator and its prey: coupled isothermal amplification of nucleic acids. *Cell Chemical Biology*, 4(1), pp.25–33.

Yuste, R. and Levin, M., 2021. New Clues about the Origins of Biological Intelligence. A common solution is emerging in two different fields: developmental biology and neuroscience. *Scientific American* [Online]. Available at: <https://www.scientificamerican.com/article/new-clues-about-the-origins-of-biological-intelligence/> [visited on 10 January 2023].

Zenil, H., 2012. *A Computable Universe: Understanding Computation & Exploring Nature As Computation*. Ed. by H. Zenil. Singapore: World Scientific Publishing Company/Imperial College Press.

But seriously: what do algorithms want? Implying collective intentionalities in algorithmic relays—a distributed cognition approach

Javier Toscano
Center for Advanced Internet Studies, (CAIS), Bochum, Germany

Abstract

Describing an algorithm can provide a formalization of a specific process. However, different ways of conceptualizing algorithms foreground certain issues while obscuring others. This article attempts to define an algorithm in a broad sense as a cultural activity of key importance to make sense of socio-cognitive structures. It also attempts to develop a sharper account on the interaction between humans and tools, symbols and technologies. Rather than human or machine-centered analyses, I draw upon sociological and anthropological theories that underline social practices to propose expanding our understanding of an algorithm through the notion of 'collective intentionalities.' To make this term clear, a brief historical review is presented, followed by an argumentation on how to incorporate it in an integral perspective. The article responds to recent debates in critical algorithm studies about the significance of the term. It develops a discussion along the lines of cognitive anthropology and the cognitive sciences, therefore advancing a definition that is grounded in observed practices as well as in modeled descriptions. The benefit of this approach is that it encourages scholars to explore cognitive

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 47–76

structures via archaeologies of technological assemblages, where intentionalities play a defining role in understanding socio-structured practices and cognitive ecologies.

Keywords

algorithm studies, distributed cognition, collective intentionalities, socio-computing infrastructures, cognitive anthropology.

Initial Definitional Attempts

Oversimplified definitions of an algorithm are currently available and frequently used, but an algorithm is neither a recipe nor a rigidly constrained and procedural formulation. Limited conceptions that represent them as a sort of entity or thing, a series of steps that need to be applied, or a simple technique that homogenizes a process, lead to weak understandings of the deeper processes, transactions and dynamics that are at stake. Indeed, an algorithm can be a problem-solving device, and this feature in itself can become a point of entry to a more complex analysis. After all, for engineers and computer scientists, "an algorithm is an abstract, formalized description of a computational procedure" (Dourish, 2016, p.3). But even if sleek and apparently elegant, the problem with this definitional reduction is twofold.

On the one hand, it concentrates on processes that happen inside computational machines. This makes the description not only machine-centered, but also introduces a misunderstanding. After all,

articulating a notion of code in the early days of computing history, pioneer logicians Newell, Simon and Shaw wrote in a seminal paper that

> the appropriate way to describe a piece of problem solving behavior is in terms of a program [. . .]. Computers come into the picture only because they can, by appropriate programming, be induced to execute the same sequences of information processes that humans execute when they are solving problems (Newell, Simon and Shaw, 1958, p.151).

In this sense, as programs, algorithms need not be thought of merely as machine drivers. And as we will see, getting rid of this idiosyncratic constraint would allow us to spot algorithms everywhere, as cultural artifacts (Finn, 2017, p.15; Seaver, 2017).

On the other hand, it is also helpful to recall that many current and historical algorithms have not implied as part of their problem-solving process to attain their objective in a neat, simple and efficient form. As a matter of fact, some algorithms aim only at keeping a solution in tension, without giving it away for everyone at every time (puzzles or riddles), while others simply produce contemplative outcomes, or even explicit nonsense (some art or literary pieces in the tradition of Dadaism, for instance). Solutions need not only be effective and efficient, they can also be creative, entertaining, experimental, contestatory, convoluted, tortuous or even purposely enigmatic or mistaken. Without reflecting on the diversity of possible outcomes and their consequences, computer scientists have usually produced equivalences between physical realities and formal symbolic systems, which have minimized the variety of "solutions" of the human world. Obviously this is not "wrong". It is what is expected from our computational machines under the dominant social normativity. But our understanding should not mistake an effect for a cause. And of course, this does not

bring us closer to a precise definition of an algorithm in the broader sense that is implied here, even if it makes clearer the scope of the task.

If we take these initial considerations into account, we can see that, in order to look for a definition of an algorithm that describes both what engineers do when they program a computer, and what users do when they tinker and apply that program, or simply invent parallel procedures for the problem they attempt to solve through any other physical technology, we need to take a different approach. First of all, we need to recognize that an algorithm is an attempt to bring something into the material world (an idea, a calculation, a previous experience). Clearly, this does not mean that every symbol will have a physical manifestation, but that symbols are intermediaries, pieces that attempt to make a translation between the ideal and the material. For as Lev Vygotsky (1978) explains, what we conventionally call tools and what we conventionally call symbols are two aspects of the same phenomenon. According to him, mediation through tools could be seen as more outwardly oriented, while mediation through signs could be seen as more inwardly oriented, toward "the self", but both aspects emerge in every cultural artifact.

If we apply this notion of artifact mediation to our search, we can see that, whether through direct tools or indirect symbols, an algorithm implies an iterative interaction with technology, or in other words, a practice of recursive intertwining between humans and the technologies they produce. Yet, for Vygotsky, an interaction with symbols or tools is not simply functional, in the sense in which a subject manipulates an object at will to achieve a task. Instead, cultural artifacts regulate interactions with one's environment and with oneself. But this is not innocuous: cultural mediation has a recursive, bidirectional effect; mediated activity simultaneously modifies both

the environment and the subject. Cultural mediation influences behaviors, synthetizes experiences from forbearers, and prepares children to acquire specific sets of accumulated memories, as knowledge (see here also Connerton, 1989). This co-constitution is what Malafouris, along a series of cognitive examinations, has termed as metaplasticity (Malafouris, 2010; 2013; 2015).

In the end, cultural mediation—or the ability to think and operate through cultural artifacts—produces historical modes of thinking—i.e. ideologies—and styles of cognition that affect how we learn, think and represent our environment and ourselves. This is what can be described as the notion of distributed cognition (Cole and Engeström, 1993; Gallagher, 2005; 2013). Laland et al. (2000, p.177) provide a definition of this process: "Distributed cognition means more than that cognitive processes are socially distributed across the members of a group. It is a broader conception that includes phenomena that emerge in social interactions as well as interactions between people and structure in their environments." The notion of distributed cognition has been a common hypothesis in linguistics and psychology ever since the writings of Vygotsky were published and made available in different translations since the 1970s, but it is by no means the standard model. There are many critics that maintain that, even if aided through different tools, thinking happens basically inside the brain (Adams and Aizawa, 2008; 2010; Loh and Kanai, 2016), or they present situations in which thinking is affected from "outside" factors (which Clowes (2019) terms as "the impact thesis"). We will not deal here with those arguments, since they appear to have a strong need for an essentialist form of conceptualizing cognition.

Moreover, since culture is for any notion of distributed cognition a foundational concept, anthropologists have made major contributions to our understanding of both the implementation of culturally

mediated forms of cognition and the various ways in which the heterogeneity of culture supports and requires the distribution of cognition. One of these anthropologists was Clifford Geertz. For Geertz, individuals submit themselves to governance by symbolically mediated programs for producing artifacts, organizing social life, or expressing emotions. In this recurring process that reaches every layer of an individual's life, humankind determines, if unwittingly, "the culminating states of its own biological destiny" (Geertz, 1973, p.48). He states, in a formulation that evokes the Vygotskian approach:

> [S]ymbols are thus not mere expressions, instrumentalities, or correlates of our biological, psychological, and social existence; they are prerequisites of it. Without men, no culture, certainly; but equally, and more significantly, without culture, no men (1973, p.49).

Geertz's formulation found strong empirical evidence among theoretical biologists, for whom the connection between culture and biology implied more than a simple correlation. As Laland et al. (2000, p.131) later would claim: "cultural traits, such as the use of tools, weapons, fire, cooking, symbols, language, and trade, may have played important roles in driving hominid evolution in general and the evolution of the human brain in particular" (see also Dunbar, 1993; or Aiello and Wheeler, 1995). Nonetheless, when Geertz and other social scientists started confining everything under the domain of "culture", throughout the 1980s, the concept became too broad and lost its specific, explanatory power. As Nick Seaver (2017, p.4) writes: "Its implicit holism and homogenizing, essentialist tendencies seemed politically problematic and ill suited to the conflictual, changing shape of everyday life." As a response, one of the most resourceful attempts in the social sciences to overcome the difficulties brought about by an

all-encompassing concept—which was nonetheless useful as a theoretical compass on a structural level—was to turn to the study of practices and symbolic interactions (Bourdieu, 1972; Certeau, 1984; Blumer, 1986). Consequently, many sociologists and anthropologists turned from a vision of a frame culture as a unified domain, to the multiplication of sites and cultures, where they could study and map emerging symbolic orders, sometimes coordinated, sometimes conflicting, out of which to make sense of the different layers of social life. This approach left behind the deterministic tone of previous explanations, with their emphasis on rules, models and texts, and began focusing instead on strategies, interests, improvisations and interactional occurrences. Recovering this emphasis, and back to our line of inquiry, Seaver (2017, p.5) provides a description of an algorithm that is worth mentioning:

> Like other aspects of culture, algorithms are enacted by practices which do not heed a strong distinction between technical and non-technical concerns, but rather blend them together. In this view, algorithms are not singular technical objects that enter into many different cultural interactions, but are rather unstable objects, culturally enacted by the practices people use to engage with them.

Seaver highlights the relational aspect of processes, enacted by practices rooted in cultural codes, therefore avoiding both a subject-centered perspective as well as a machine-centered view. In that sense, his definition is in line with a number of interesting theories and methodologies that have emerged in sociology and science and technology studies over the past two decades, for example: actor-networks (Callon, 1986; Latour, 1992; 2005), sociotechnical ensembles (Bijker, 1999), object-centered socialities (Knorr Cetina, 1997), relational materialities (Law, 2004), constitutive entanglements (Orlikowski, 2007)

or object-oriented ontology (Harman, 2002; Bryant, 2010), as well as the approach of cognitive ecology (Hutchins, 2010) and material engagement theory (Malafouris, 2005; 2013) in the cognitive sciences. These theories challenge and transcend conventional distinctions between objects and subjects, as well as between social abstractions and material iterations. Furthermore, their particular value lies in their insistence on speaking of the social (e.g. culture) and the material (e.g. nature) in the same register, and on not resorting to a limiting dualism that treats them as separate, even if interacting, phenomena.

Being an anthropologist, Seaver concentrates on the instabilities, the discontinuities, the confusions, the contradictions and the misunderstandings that enable different traditions and enrich human life. However, his view can be further explored, since it lacks a reasonable explanation of how, despite being categorized as "unstable objects", algorithms may appear as robust, reliable and even intrinsically repeatable. In other words, how do procedural patterns are sustained, despite variance; how consistencies emerge to enable traditions; when are recurrences broken up and when are they maintained? These inquiries are relevant because algorithms are something more than people executing socially available recipes and tweaking them with a personal taste. Algorithms are clusters of affordances and patterns that emerge in every process of *recursive* intertwining between humans and technologies. In that sense, they could be seen as material or immaterial scripts that link mental states with both material procedures and technocultural resources, enacted as a cultural practice to accomplish a specific task (effectively or not). And yet, in this description, mental states need not pertain to a single individual. Actually, if they would really belong to a unique individual (someone looking for a unique solution to his/her own problems, desires or needs) they would be

socially illegible. But shouldn't this call for the inference of collective mental states? And what would that entail? The issue demands a deeper inspection, and we will now turn to it.

The mental and the notion of collective intentionalities

In order to inspect closer how humans and technologies interact through material or immaterial procedures linking mental states to real-world conditions, we need to acknowledge what we mean by mental states, and how they emerge as techno-cultural practices out of which specific patterns can be traced. This will require a short detour to explain some basic conceptions, but by the end of this explanation we will have a clearer landscape of the categories at stake.

A mental state can best be delineated by the notion of intentionality. Intentionality is a complex philosophical concept that emerged with Medieval Scholasticism through Medieval Islamic philosophy, but was later retaken and developed in phenomenological circles, starting from the 19th century. Franz Brentano's work is usually set as a point of departure for contemporary analyses. In his writings, intentionality is set as an attribute of an individual's mind, which adheres to mental contents, as opposed to attributes of the real world, such as extension and duration, which can be predicated of existing objects. Brentano takes on the discussion from St. Thomas Aquinas, who established that the object which is thought is intentionally in the thinking subject, the object which is loved in the person who loves, the object which is desired in the person desiring, etc. In that sense,

intentionality is clearly something that can be predicated of inexistent phenomena, but which has an effect on our own conceptions, desires and beliefs. Brentano (1995, p.68) writes:

> Every mental phenomenon is characterized by [...] the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood here as meaning a thing), or immanent objectivity. Every mental phenomenon includes something as object within itself [...] We can, therefore, define mental phenomena by saying that they are those phenomena which contain an object intentionally within themselves.

At this point, intentionality was described as a clear attribute of mental activity, independent of a real world, but clearly related to it, and decisive to ascribe it meaning. This trait was important because it offered a form of cognizing reality without relying on the Kantian formulation that attempted to align (individual) sensations and (social) concepts. In other words, it created a model where things could be cognized beyond a thick web of structured epistemological pre-conceptions. This is precisely what encouraged Husserl's enthusiasm, as inscribed in his motto "Back to the things themselves!" (*Zurück zu den Sachen selbst!*). For as Merleau-Ponty (2005, p.xix) writes:

> What distinguishes intentionality from the Kantian relation to a possible object is that the unity of the world, before being posited by knowledge in a specific act of identification, is 'lived' as ready-made or already there.

However, intentionality in this early stage also made a clear difference between the inner, mental world, and the outer, objective reality. In that sense, it was still trapped in the fundamental dualism

that characterized the positivist style of thinking in the late 19th and early 20th centuries. This dynamic has been sufficiently deconstructed, especially within the theories that were mentioned in the previous section, and there is no need to discuss it further. A second problem is that this early notion of intentionality also posited a very clearly delimited "self" for whom an intention (and communication of that intention) is transparent. The precise refutation of this point can be extensive, and it can also run through diverging lines, but for synthetic aims, we can resort back to the Vygotskian approach and understand the "self" as a symbol and a cultural artifact. Actually, both Vygotsky and a contemporary anthropologist of him, G.H. Mead, worked along the lines of a similar hypothesis, which has been termed the "social genesis of the self" (Glock, 1986), and which implied both the process of internalisation (through education in the child) and the genesis of linguistic meaning. For Mead (1972, p.164), for instance, "[t]he process out of which the self arises is a social process which implies interaction of individuals in the group, implies the preexistence of the group." Accordingly, he adds:

> the self appears in experience essentially as a "me" with the organization of the community to which it belongs. This organization is, of course, expressed in the particular endowment and particular social situation of the individual [...]. He is what he is in so far as he is a member of this community, and the raw materials out of which this particular individual is born would not be a self but for his relationship to others in the community of which he is a part (Mead, 1972, p.200).

Following the Vygotsky/Mead hypothesis, there cannot even be a "direct" connection between an individual and her experience, because this connection is mediated through language, by which a "self" appears as some type of thing. In other words, the emergence of a "self"

is an effect, or a functional construction, of a subject that has learned how to enunciate and use the particle "I" under a given set of socially-sanctioned, grammatical rules. This brings us to a rather interesting situation on the cognitive side. For if the self is a social construction, what is to be done with what we call "the mental"? Is the link between both notions merely a deficient attribution, or is it a faulty causal connection? Mead describes the mental as an emergent phenomenon, which involves a relationship to the character of things:

> Those characters are in the things, and while the stimuli call out the response which is in one sense present in the organism, the responses are to things out there. The whole process is not a mental product and you cannot put it inside of the brain. Mentality is that relationship of the organism to the situation which is mediated by sets of symbols (Mead, 1972, pp.124–125).

This turns irrelevant the attribution of mentality to the self. On the same grounds, a causal connection between them can only be inferred as inexistent. Instead, both are equally emergent effects of a given symbolic mediation. Mead's description of the mental (that cognitive relationship of an organism to a situation, mediated by symbols) is the backbone to the definition of an algorithm that was proposed on the previous section. It is also a touchstone in the tradition of cognitive anthropology that has been associated with the idea of cognitive ecologies (Douglas, 1986; Lave, 1988; Connerton, 1989; Hutchins, 1995; 2010), as well as in traditions of cognitive sciences that inquire into models of an embodied, embedded, extended and/or an enactive social mind (Clark, 1997; 2003; 2015; Clark and Chalmers, 1998; Gallagher, 2005; 2013; Gallagher and Miyahara, 2012). Gallagher (2013, p.4), for instance, describes the mental in this way:

> If we think of the mind not as a repository of propositional attitudes and information, or in terms of internal belief-desire psychology, but as a dynamic process involved in solving problems and controlling behavior and action—in dialectical, transformative relations with the environment—then we extend our cognitive reach by engaging with tools, technologies, but also with institutions. We create these institutions via our own (shared) mental processes, or we inherit them as products constituted in mental processes already accomplished by others.

Indeed, breaking the causal link between the mind and the self allow us to see the dense and emergent network of affordances and enactions that constitute cognitive phenomena. But how do intentionalities come back into the picture? For Brentano, intentionalities were so much as the mark of the mental, i.e. the defining quality of an inexistent, psychological phenomenon. But if the mind is not any more located in an inner, private world, should we just simply do without them? Quite the opposite. As a matter of fact, intentionalities play a stronger role within a distributive cognition approach. But we need to refine the conceptual frame to see how this can be integrated into a comprehensive explanation.

An intentionality is not a purpose, nor a design or an intention to do something, although the notions are closely related. Actions are intentional, for example, not only because there is a will behind them, but also because they follow a goal or a project. If I am hungry and I do not have anything to eat at home, I can go out to a supermarket to buy groceries in order to cook, or to a restaurant, or even to a place where food is distributed if my economic means are limited. These, among others, are available modes of action, connected to material and technical functions, social behaviors and actionable symbolic networks. But we know that there used to be a time when, if hun-

gry, people could go out hunting or foraging, and the relevant social programs were there to support those activities. Intentionalities are attached then to historical norms, cultural repertoires, social habits, communal values, rituals and many other forms and forces that can be seen to shape an individual's action. For as Brandom (1994, p.61) writes:

> only communities, not individuals, can be interpreted as having an original intentionality. [T]he practices that institute the sort of normative status characteristic of intentional states must be social practices.

In that sense, the social life of an individual consists in a good deal in determining the appropriateness of her own desires and needs as these are articulated to the available social practices, or cultural programs, through inferential reasoning, practical adjustments and other means.

Now, even if at a first look this explanation seems to restrain an individual's agency, by making her guide a certain "intended" action through a given catalogue of socially sanctioned paths, the picture that this model enables is in fact richer and more complex. In a few words, a strict functionalism does not apply (Elster, 1983; Douglas, 1986, p.32ff). As a matter of fact, a model like Malafouris' material engagement theory actually sustains that the distinctive forms of human agency emerge precisely in the practical space afforded by the interactions (Malafouris, 2008; 2015). After all, an individual never "acts" in a void either. And as Cooren et al (2006, p.11) write:

> Agency is not a 'capacity to act' to be defined a priori. On the contrary, it is 'the capacity to act' that is discovered when studying how worlds become constructed in a certain way.

In that sense, intentionalities are sustained in social practices without losing their capacity for an individual's adaptation, expression and

further innovation. And as such, they can be acknowledged as *collective intentionalities*, fundamental pieces that connect an individual to a larger collective, without necessarily turning them into a deterministic setup. Collective intentionalities are in that sense something as action-able paths, through which an individual orients and articulates her actions with the resources and experiences of a cultural community, i.e. a community of practice.

Furthermore, collective intentionalities are so relevant that Tomasello (2014) assigns to them, in an appealing hypothesis, a definite role in the evolution of the species, since they allow coordination and cooperation to occur not only simultaneously, but also throughout generations. For this cognitive linguist, collective intentionalities comprise

> not just symbolic and perspectival representations but conventional and 'objective' representations; not just recursive inferences but self-reflective and reasoned inferences; and not just second-personal self-monitoring but normative self-governance based on the culture's norms of rationality (Tomasello, 2014, p.6).

As such, they are the infrastructure of social life, underlying even culture and language through pre-linguistic aims and forces that acquire a given shape. Developing over the foundations of collective intentionalities,

> culture and language, as agent-neutral conventional phenomena [...] provide another setting within which a new form of human sociality can lead to a new form of human thinking, specifically, objective reflective-normative thinking (Tomasello, 2014, p.141).

In that sense, collective intentionalities can be said to be the building blocks of human-symbol/tool interactions. But in the end, if collective

intentionalities are not a quality of the objective world—but rather its foundation—where are these to be seen, or how do they emerge and provide tangible samples for interactions to occur? We will tackle the issue in the following section.

Collective intentionalities and algorithms: heuristics and dynamics

The notion of collective intentionality is only such if it retains one condition that was there since Brentano attempted a definition: it is an attribute of a mental state, i.e. a mark of the mental. But we have seen that, in a distributed cognition approach, the mental cannot be exclusively associated with a self; it is rather an articulated web that links individuals to tools and symbols that have been pre-structured by a collective, and are enacted through social practices. So we are presented with an empirical challenge: how to spot an intentionality if it is neither an objective nor a subjective phenomenon in the classical sense? Collective intentionalities are usually "hidden" to the naked eye, sometimes they are by-products of repeated actions, much as a trailing path in the woods which appears after years and years of different individuals walking through it, but sometimes they also stand out in oblique moves. In any case, they comprise the causal loops that run behind collective articulations (making up a good deal of group identities, for example), and these can emerge as latencies, background or naturalized conventions (a specialized analysis in Chant, Hindriks and Preyer, 2014; in relation to this topic see also Toscano, forthcoming). The only minimal assumption is that they stand in a threshold, as that which allows community survival without demanding from

individuals that they give up on their autonomy (even though the threshold is dynamic, and is not the same for a child as for the elder, or throughout different knowledge capacities and hierarchies).

In the last years, cognitive scientists have developed different models to locate intentionalities via distinctive approaches. The neo-behaviourist Daniel Dennett, for instance, ascribes intentionality to observed rational behavior, and he describes the agent as someone

> who harbors beliefs and desires and other mental states that exhibit intentionality or 'aboutness', and whose actions can be explained (or predicted) on the basis of the content of these states (Dennett, 1991, p.76).

The approach is clear, and amounts to correlating traces to directions and motivations in a straightforward way. Of course, it is constrained to reading rational behavior and to valuing every action as instrumental to achieve a specific goal. In contrast, a neo-pragmatist view (Brandom, 1994; 2000; Cash, 2008; 2009) proceeds by ascribing intentionality as an explanation and a specific coupling of action to social norms. As Cash (2008, p.101) argues:

> based on the similarity of their movement to the kind of actions, [. . .] would entitle us to ascribe such intentional states as reasons.

This might be a key aspect in certain contexts, but it is constrained to knowing what the norms to be applied are, and to evaluating if the ensuing pairing of actions to those norms succeed or not. In that sense, they imply the recognition of patterns, and a judgment on their application or continuity, but they also underestimate the value of deviance and disregard a space for individual creativity. Even a third approach, which we can call a neo-interactionist perspective, aims at understanding other's intentionalities not by acknowledging

or judging their actions, but by understanding actual or potential interactions with others in socially appropriate ways. As Gallagher and Miyahara (2012, p.135) write in this account:

> we normally perceive another's intentionality in terms of its appropriateness, it's pragmatic and/or emotional value for our particular way of being, constituted by the particular goals or projects we have at the time, or implicit grasp on cultural norms, our social status, and so on, rather than as reflecting inner mental states, or as constituting explanatory reasons for her further thoughts and actions.

The neo-interactionist perspective certainly rounds up some of the forms in which intentionalities emerge, but they do not completely revoke the previous explanations, and instead helps compile a catalogue of intentional enactions.

At this point, we can bring back the definition of an algorithm that was proposed in the first section, and mobilize it in an illustrative form. We can thus define an algorithm as

> a recursive script that links collective intentionalities with both material procedures and technocultural resources, enacted as a cultural practice to accomplish a specific task.

We now know what is meant by a collective intentionality. And we can expect how to look for them. But this definition does more than just describe a process. It wants to reflect on the fact that collective intentionalities are not by themselves the structures that sustain a community's culture. It is really their mobilization, in an algorithmic form, which brings them to life. It would therefore be more precise to see an algorithm as an action than as an object, however "unstable" that object would turn out to be. On those grounds, an algorithm should be

seen more as an activity, an "algorithmation", a productive emergent pattern that enables connections out of a given networked system or a distinctive cognitive ecology.

Within the algorithmic feedback loops, the individual performs a key computational function. Clearly, since we do not rely on a machine-centered perspective, a corresponding idea of computation must be outlined. For simplicity, we can take over Hutchins view here. He suggests that computation should be regarded as "the propagation of representational states across representational media" (Hutchins, 1995, p.118). In that sense, individuals are the agents transforming representational states for those collective intentionalities through algorithmic procedures, that is, through recursive technical enactions. But in this finely threaded network, the individual is neither the origin nor the final end. And yet, she is not a simple cog in the system either. She is interconnected, interacting, adjusting herself and her environment with this complex and finely tuned mechanism, which we might indeed call at this point a socio-computing infrastructure (Toscano, forthcoming). Yet this term cannot imply a fix and immovable architecture, but a dynamic structure where certain accomplishments, and not others, are viable. Laland et al. (2000, p.130), for instance, refer as "niche construction" to the human-made or human customized structures that are essential to the development, production and continuance of certain activities. Jones et al. (1997) identify that same activity as "ecosystem engineering". The notion of socio-computing infrastructure that is proposed here here should be read along those lines, but where the accent on collective intentionalities and a social cognitive activity is deliberate.

Similarly, Laland et al. (2000) propose the idea of an ecological niche, which implies that an organism occupies a distinctive role in each ecosystem. This opens up yet another approach to the task of

identifying intentionalities, as a supplement to the ones that were described above. For in certain contexts, defining the ecological niche of an individual can render a better inspection of the collective intentionalities implied in a given system. In other words, in human-machine interactional systems, focusing on the active or operating roles of given individuals as socially-enabled subjective behaviors or social functions can shed light on the specific collective intentionalities at work. This route acknowledges that the individual is relevant, only not on her own, but through her dynamic links (interpretations, associations, appreciations) to a broader community of practice. This can be useful in anthropological cases, but is doubtlessly crucial in historical inquiries and techno-archeological analyses. We can bring a couple of empirical cases from this latter for consideration.

a) Inka's Khipus

If we think of historical socio-computing infrastructures that were lost or disrupted when the groups tied to them ceased to exist, we can acknowledge which were precisely the missing access points that make the reconstruction or re-interpretation of those systems difficult, or sometimes impossible. Two cases can be explored here at some length. As a first case, we can recall the recent decipherment of ancient *khipus* in Peru. Khipus were devices of statistical notation that stemmed during the Inka Empire, but were used until the Spanish colonial period in that South American country. These devices did not employ numerical symbols, but relied instead on cotton strings of different lengths and colors, and were encoded using knots at different places. As Medrano and Urton (2018, p.2) state:

> the Inkas filled the twists and knots of the khipus with data, including bureaucratic accounting measures such as tax assignments and census counts.

One element is noticeable when approaching the khipu coded system: the people that used it as a statistical artifact did not suddenly disappeared without leaving a trace (as the Maya civilization did, for instance). On the contrary, the khipus coexisted during some time with the European statistical methods of the epoch, which the Spanish had brought with them. In that situation, even if symbolic and abstract operations were readily available, khipus were kept because they implied a material manifestation of different social values, symbols that the people of that particular culture considered relevant information, as opposed to mere abstractions. In other words, khipus enriched merely numerical data: they registered social relations of a highly organic and interpersonal nature, traits that were indifferent to the Spanish accounting methods, which were therefore inadequate for their transmission (Medrano and Urton, 2018, p.12).

In the Inka worldview, khipus were not only statistics, but a representation of a given reality made possible through a material craft. Of course, since the symbols they employed were not easily manageable, the khipus were discontinued after some time. Nobody wrote how they were encrypted, so the key to reading them disappeared. In a sense, khipus were meant not only as notational systems, but also as mnemonic devices for khipu keepers and scribes. When these professionals finally changed the notational system to make their calculations, the mnemonic function ceased to operate. But while still active, these professionals were implementing an algorithmic procedure: they applied a know-how for a given collective intentionality—to count, or calculate, a given state of affairs—and turned it into an objectified device—a social representation—thus computing it. The khipus were finally deciphered through an analogy with an European-style census that was later discovered to match one of these objects with a strict correlation, but also by paying a close attention to Inka's testimonies

on economy, politics, religion and other aspects of their civilization that were highly valued, and considered to be worthy of a specific notational foundation.

b) The Voynich Manuscript

Another case is provided by the situation of the *Voynich Manuscript*, kept in the collection of rare books at the Beinecke Library, at Yale. This fifteenth-century codex has not been deciphered until this day for several reasons, many of which are elements that indicate how a socio-computational infrastructure, and with it a specific algorithmic enaction, is put to work. The "book" was written in an unknown script by an unknown author. The impossibility to assign it a context, a precise culture, or even a specific function within a given literary or scientific biography, contribute to see this piece as an example of a radical particularity that highlights its isolatory character. This is just not how a "book" works. Rather than executing a typical communicative intentionality, the *Voynich Manuscript* contradicts its form and function, and appears as a work of madness. The current custodians of the book present it thus: "the manuscript has no clearer purpose now than when it was rediscovered in 1912" (Clemens, 2016). There are no points of access because nobody knows where to begin with. Of course, some facts can be determined: the approximate date of its physical appearance, as well as a list of its owners, all of which tempt the researchers to make some claims based on analogical and normative assumptions, of the kind that cognitive scientist have shown how to bring about. But in the end, the manuscript has been annulled as an informational device, as well as an instrument of contextual cognition. However, it has become a new source of computations, for the curiosity of the researchers has turned it into an object of study,

which means that it is being transfigured across different representational media. In any case, without an anchoring fact that stabilizes its meaning, such investigations speak more about our computational procedures than about the content of the "book", so they also tell about our need for conceptual pre-assumptions and our own inabilities to understand even human-made objects when a clear intentionality is not recognized or set onto them.

Conclusions

Algorithms cannot be reductively described as machine drivers or mere coding language. They imply instead a complex cultural activity that involve both material and immaterial interactions. This article has aimed to show how, as part of their particular enaction, they are constructed along collective intentionalities of different sorts. In that sense, algorithms do shape desires, wants and needs, as these are ingrained in distinctive communities. It is indeed through an algorithmic recursiveness that collective cultures flourish and expand. It is also through an individual's tinkering with them that they can give way to adjustments and innovations, provided that the underlying intentionalities—whether as paths, patterns, occurrences or scripts—remain fundamentally recognizable.

In his book *What do Algorithms want*?, Ed Finn finds an ingenuous answer to this complex question: "This is what algorithms want, or what we design them to want: to know us completely" (Finn, 2017, p.82). But this statement is a simplification that requires further clarification itself. Algorithms cannot want something in themselves, but neither do we. Or the other way around: algorithms want what "we" want, or rather: we want through them. Which is not always something

evident. After all, "to want" is a cultured habit, which is ingrained in children through upbringing and education. As individuals, we use socially available algorithms to channel pre-linguistic and abstract desires and needs, which only through them acquire a definable form. So in a way it is true: algorithms want what we design them to want. But we can only design what is culturally available, collectively interpretable, socially desirable. So it is less true that we design all algorithms "to know us completely". In fact, most of the time, the opposite is just the case. In their recursivity, algorithms enact collective intentionalities that are frequently turned into latencies, background or naturalized conventions, and then cease to appear as constructions to us. (Therefore, only in a culture where information extraction is a viable practice, the design of algorithms to extract information from us—what Finn refers as "to know"—will be a logical consequence.) In the end, algorithms imply an articulatory activity: they are collective processes of cultural inscription, through which individuals enact socially available programmatic technologies for a specific, intentional objective.

This article has sought to provide examples on how to approach collective intentionalities, both by recalling how cognitive scientists apply logical inferences to distinguish emergent phenomena, and by turning to historical socio-computing infrastructures to inspect their legibility (or lack thereof) and operation. Evidently, much works needs to be done to deepen a techno-archeological inquiry of this kind, but this article has sought to contribute with some entry points to enrich such analyses in a distinctive way.

Bibliography

Adams, F. and Aizawa, K., 2008. *The Bounds of Cognition*. 1st ed. Malden, MA: Blackwell Publishing.

Adams, F. and Aizawa, K., 2010. Defending the Bounds of Cognition. In: R. Menary, ed. *The Extended Mind* [Online]. MIT Press, pp.67–80. https://doi.org/10.7551/mitpress/9780262014038.003.0004.

Aiello, L.C. and Wheeler, P., 1995. The Expensive-Tissue Hypothesis: The Brain and the Digestive System in Human and Primate Evolution. *Current Anthropology* [Online], 36(2), pp.199–221. https://doi.org/10.1086/204350.

Bijker, W.E., 1999. *Of Bicycles, Bakelites, and Bulbs: Toward a Theory of Sociotechnical Change*. 3rd ed., *Inside technology*. Cambridge, MA: MIT Press.

Blumer, H., 1986. *Symbolic Interactionism: Perspective and Method*. Berkeley: University of California Press.

Bourdieu, P., 1972. *Outline of a Theory of Practice*. Cambridge, MA: Cambridge University Press.

Brandom, R.B., 1994. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.

Brandom, R.B., 2000. *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.

Brentano, F., 1995. *Psychology from an Empirical Standpoint [1874]*. Ed. by O. Kraus and L.L. McAlister (A.C. Rancurello, D. Terrell and L.L. McAlister, Trans.), *International Library of Philosophy*. London; New York: Routledge.

Bryant, L.R., 2010. *Onticology–A Manifesto for Object-Oriented Ontology Part I*. Available at: <https://larvalsubjects.wordpress.com/2010/01/12/object-oriented-ontology-a-manifesto-part-i/> [visited on 18 January 2023].

Callon, M., 1986. Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of Saint Brieuc Bay. In: J. Law,

ed. *Power, action, and belief: a new sociology of knowledge?*, *Sociological review monograph*, 32. London; Boston: Routledge & Kegan Paul, pp.196–233.

Cash, M., 2008. Thoughts and oughts. *Philosophical Explorations* [Online], 11(2), pp.93–119. https://doi.org/10.1080/13869790802015635.

Cash, M., 2009. Normativity is the mother of intention: Wittgenstein, normative practices and neurological representations. *New Ideas in Psychology* [Online], 27(2), pp.133–147. https://doi.org/10.1016/j.newideapsych.2008.04.010.

Certeau, M.d., 1984. *The Practice of Everyday Life* (S. Rendall, Trans.). Berkeley; Los Angeles, CA: University of California Press.

Chant, S.R., Hindriks, F. and Preyer, a.G., eds., 2014. *From Individual to Collective Intentionality: New Essays*. Oxford; New York: Oxford University Press.

Clark, A., 1997. *Being There: Putting Brain, Body, and World Together Again*. Cambridge; London: MIT Press.

Clark, A., 2003. *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford; New York: Oxford University Press.

Clark, A., 2015. What 'Extended Me' knows. *Synthese* [Online], 192(11), pp.3757–3775. https://doi.org/10.1007/s11229-015-0719-z.

Clark, A. and Chalmers, D., 1998. The extended mind. *Analysis* [Online], 58(1), pp.7–19. Available at: <https://www.jstor.org/stable/3328150> [visited on 3 June 2019].

Clemens, R., ed., 2016. *The Voynich Manuscript*. New Haven; London: Yale University Press.

Clowes, R.W., 2019. Screen reading and the creation of new cognitive ecologies. *AI & Society* [Online], 34(4), pp.705–720. https://doi.org/10.1007/s00146-017-0785-5.

Cole, M. and Engeström, Y., 1993. A Cultural-Historical Approach to Distributed Cognition. In: G. Salomon, ed. *Distributed Cognitions: Psychological and Educational Considerations*, *Learning in doing*. Cambridge, MA: Cambridge University Press, pp.1–46.

Connerton, P., 1989. *How Societies Remember* [Online]. 1st ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511628061.

Cooren, F., Taylor, J.R. and Van Every, E.J., 2006. *Communication as Organizing: Empirical and Theoretical Approaches into the Dynamic of Text and Conversation*, *LEA's communication series*. Mahwah, NJ: Lawrence Erlbaum.

Dennett, D.C., 1991. *Consciousness Explained*. Boston: Little, Brown and Company.

Douglas, M., 1986. *How Institutions Think*. 1st ed., *The Frank W. Abrams lectures*. Syracuse, NY: Syracuse University Press.

Dourish, P., 2016. Algorithms and their others: Algorithmic culture in context. *Big Data & Society* [Online], 3(2), pp.1–11. https://doi.org/10.1177/2053951716665128.

Dunbar, R.I.M., 1993. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences* [Online], 16(4), pp.681–694. https://doi.org/10.1017/S0140525X00032325.

Elster, J., 1983. *Explaining Technical Change: A Case Study in the Philosophy of Science* [Online], *Studies in Rationality and Social Change*. Cambridge; London et al.: Cambridge University Press. Available at: <http://archive.org/details/Elster1983ExplainingTechnicalChangeACaseStudyInThePhilosophyOfScienceBook> [visited on 18 January 2023].

Finn, E., 2017. *What Algorithms Want: Imagination in the Age of Computing* [Online]. 1st ed. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/10766.001.0001.

Gallagher, S., 2005. *How the Body Shapes the Mind*. Oxford; New York: Clarendon Press.

Gallagher, S., 2013. The socially extended mind. *Cognitive Systems Research* [Online], 25-26, pp.4–12. https://doi.org/10.1016/j.cogsys.2013.03.008.

Gallagher, S. and Miyahara, K., 2012. Neo-Pragmatism and Enactive Intentionality. In: J. Schulkin, ed. *Action, Perception and the Brain: Adaptation and Cephalic Expression* [Online], *New Directions in Philosophy and Cognitive Science*. London: Palgrave Macmillan UK, pp.117–146. https://doi.org/10.1057/9780230360792_6.

Geertz, C., 1973. *The Interpretation of Cultures: Selected Essays*. New York: Basic Books.

Glock, H.-J., 1986. Vygotsky and Mead on the Self, Meaning and Internalisation. *Studies in Soviet Thought* [Online], 31(2), pp.131–148. Available at: <https://www.jstor.org/stable/20100086> [visited on 18 January 2023].

Harman, G., 2002. *Tool-Being: Heidegger and the Metaphysics of Objects*. 1st ed. Chicago: Open Court.

Hutchins, E., 1995. *Cognition in the Wild*. Cambridge, MA; London: MIT Press.

Hutchins, E., 2010. Cognitive ecology. *Topics in Cognitive Science* [Online], 2(4), pp.705–715. https://doi.org/10.1111/j.1756-8765.2010.01089.x.

Knorr Cetina, K., 1997. Sociality with Objects: Social Relations in Postsocial Knowledge Societies. *Theory, Culture & Society* [Online], 14(4), pp.1–30. https://doi.org/10.1177/026327697014004001.

Laland, K.N., Odling-Smee, J. and Feldman, M.W., 2000. Niche construction, biological evolution, and cultural change. *Behavioral and Brain Sciences* [Online], 23(1), pp.131–146. https://doi.org/10.1017/S0140525X00002417.

Latour, B., 1992. Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts. In: W.E. Bijker and J. Law, eds. *Shaping technology/building society: studies in sociotechnical change*, *Inside technology*. Cambridge, MA: MIT Press, pp.225–258.

Latour, B., 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*, *Clarendon lectures in management studies*. Oxford; New York: Oxford University Press.

Lave, J., 1988. *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life* [Online]. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511609268.

Law, J., 2004. *After Method: Mess in Social Science Research*, *International library of sociology*. London; New York: Routledge.

Loh, K.K. and Kanai, R., 2016. How Has the Internet Reshaped Human Cognition? *The Neuroscientist* [Online], 22(5), pp.506–520. https://doi.org/10.1177/1073858415595005.

Malafouris, L., 2005. The Cognitive Basis of Material Engagement: Where Brain, Body and Culture Conflate. In: E. DeMarrais, C. Gosden

and C. Renfrew, eds. *Rethinking Materiality: Engagement of Mind with Material World*. Cambridge: McDonald Institute for Archaeological Research, pp.53–61.

Malafouris, L., 2008. At the Potter's Wheel: An Argument for Material Agency. In: C. Knappett and L. Malafouris, eds. *Material Agency: Towards a Non-Anthropocentric Approach* [Online]. Berlin: Springer, pp.19–36. https://doi.org/10.1007/978-0-387-74711-8_2.

Malafouris, L., 2010. Metaplasticity and the human becoming: principles of neuroarchaeology. *Journal of Anthropological Sciences*, 88(4), pp.49–72.

Malafouris, L., 2013. *How Things Shape the Mind: A Theory of Material Engagement*. Cambridge; London: MIT Press.

Malafouris, L., 2015. Metaplasticity and the Primacy of Material Engagement. *Time and Mind* [Online], 8(4), pp.351–371. https://doi.org/10.1080/1751696X.2015.1111564.

Mead, G.H., 1972. *Mind, Self and Society: From the Standpoint of a Social Behaviorist [1934]*. Ed. by C. Morris. Chicago; London: University of Chicago Press.

Medrano, M. and Urton, G., 2018. Toward the Decipherment of a Set of Mid-Colonial Khipus from the Santa Valley, Coastal Peru. *Ethnohistory* [Online], 65(1), pp.1–23. https://doi.org/10.1215/00141801-4260638.

Merleau-Ponty, M., 2005. *Phenomenology of Perception: An Introduction [1945]* (C. Smith, Trans.), *Routledge classics*. London; New York: Routledge.

Newell, A., Simon, H.A. and Shaw, J.C., 1958. Elements of a theory of human problem solving. *Psychological Review* [Online], 65(3), pp.151–166. https://doi.org/10.1037/h0048495.

Orlikowski, W.J., 2007. Sociomaterial Practices: Exploring Technology at Work. *Organization Studies* [Online], 28(9), pp.1435–1448. https://doi.org/10.1177/0170840607081138.

Seaver, N., 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* [Online], 4(2), pp.1–12. https://doi.org/10.1177/2053951717738104.

Tomasello, M., 2014. *A Natural History of Human Thinking*. Cambridge, MA: Harvard University Press.

Toscano, J., forthcoming. Intentionalities of the algorithm: historical practices and socio-computing infrastructures. A philosophical account. *Lo Sguardo*, 34(1).

Vygotsky, L.S., 1978. Tool and Symbol in Child Development. In: M. Cole, V. John-Steiner, S. Scribner and E. Souberman, eds. *Mind in Society: The Development of Higher Psychological Processes*. Cambridge: Harvard University Press, pp.19–30.

Modelling interactive computing systems: Do we have a good theory of what computers are?

Alice Martin
Mathieu Magnaudet
Stéphane Conversy
ENAC, Université de Toulouse, France

Abstract

Computers are increasingly interactive. They are no more transformational systems producing a final output after a finite execution. Instead, they continuously react in time to external events that modify the course of computing execution. While philosophers have been interested in conceptualizing computers for a long time, they seem to have paid little attention to the specificities of interactive computing. We propose to tackle this issue by surveying the literature in theoretical computer science, where one can find explicit proposals for a model of interactive computing. In that field, the formal modelling of interactive computing systems has been brought down to whether the new interaction models are reducible to Turing Machines. There are three areas where interaction models are framed. The comparison between TMs and interactive system models is at stake in all of them. These areas are namely some works on concurrency by Milner, on Reactive Turing Machines, and on interaction as a new computing paradigm. For each of the three identified models, we present its motivation, sum up its account for interaction and its legacy, and point out issues regarding the understanding of computers. The survey shows difficulties for epistemologists. The reason is that these analyses focus

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 77–119 • CC-BY-NC-ND 4.0

on the formal equivalence between interactive models of computation and classic ones. Such a project is different from addressing how a computing machine can be interactive: in other words, which mechanisms allow it.

Keywords

philosophy of computing, models of computation, interactive computing, computing mechanism, computational mechanistic explanation.

Introduction

In the philosophy of computing, we are paying increased attention to a set of new features of computers. This set has led to the introduction of a new label for these computing machines: they are referred to as interactive computing machines (Dodig-Crnkovic, 2011; Goldin, Wegner and Smolka, 2006; Soare, 2013; Van Leeuwen and Wiedermann, 2001; Wegner, 1997). The set of new features can be captured in the following statement made in a 2011 paper by Gordana Dodig-Crnkovic (our highlights):

> Present day computers are very different from the early standalone calculators designed for mechanizing mathematical operations. They are largely *used for communication in world-wide networks* and variety of information processing and knowledge management. Moreover, they play *an important role in the control of physical processes and thus connect to the physical world*, especially in automation and robotics. [...] Computational processes are nowadays *distributed, reactive, agent-based, and concurrent*. The main criterion of success of the computation is not its termination, but its *response to the*

> *outside world*, its speed, generality and flexibility; adaptability, and tolerance to noise, error, faults, and damage (Dodig-Crnkovic, 2011).

Historically, the concept of interaction was introduced by a computer scientist, Milner, in the 1970s-1980s (Milner, 1975; 1982; 1993; 1999). At first, an *interactive computing system* was defined as a system where several threads execute instructions in parallel while being able to synchronize and communicate at certain moments of the execution. Since then, the characteristics of computer systems have continued to evolve, and by "interactive" we refer today to a broader set of properties that can be grouped as follows: the ability to continuously react in time to external events that modify computing execution. This class of computers deserves all our attention since they are ubiquitous. Every computer system today is designed to respond to external events in a predictable way and according to temporal constraints. In any case, what distinguishes this class of so-called interactive computing machines from the classical computer systems that preceded them is that they are no longer purely *transformational systems*. A *transformational system* is a classical computing device that, given a set of inputs, produces a final output after a finite execution. This evolution of computing complicates the answer to *what a computer is*. The question is well-known in the philosophy of computing (Piccinini, 2008; Rapaport, 2018; Smith, 2002). As already noted, many answers to the question distort it and answer the question of *what a computation is*, immediately projecting the field of investigation into the theory of computability:

> A fairly obvious, trivial, and almost-circular definition of 'computer' says that a computer is a machine that computes. The natural next question is: What does it mean to compute? But this shifts the burden of answering our question away from

> what computers are to the topic of what computation is. Many of the objections to various theories about computers are really objections to what counts as a computation (Rapaport, 2018).

This leaves us with the specific issue we want to address. We ask whether models of computation for interaction allow us to answer the question of what an actual (necessarily interactive) computer is. Current computers come in various forms and we chose in this paper to restrict our concerns to a delimited notion of *interaction*, as defined in Human-Computer Interaction (Basman et al., 2018; Beaudouin-Lafon, 2006; Dearden and Harrison, 1997; Hornbaek and Oulasvirta, 2017; Myers, 1994), and target a specific set of ubiquitous computing devices—those interacting with humans, e.g., through digital interfaces. We will not elaborate on *analog computing* (Bielecki, 2019) and *natural computation* (Dodig-Crnkovic, 2011; MacLennan, 2003).

To tackle the issue of interactive computing devices, we propose here an approach that, to the best of our knowledge, has not been proposed so far: we want to examine the models of computation proposed in theoretical computer science to think about interactive computing systems. We offer a literature survey where one can find *explicit theories of interaction*.[1] We show that the formal modelling of interactive computing systems has been brought down to whether the new interaction models are reducible to Turing's *a-machines* (Turing, 1937)—we will refer to them as Turing Machines (TMs). Questioning the theoretical bounds of the Turing Machine in computer science when faced with the existence of interactive computing devices has been explored at least since Milner's work on communicating and

[1] We insist on our two criteria: *explicit theories of interaction* in *theoretical computer science*. We have in mind the fact that other communities e.g., the engineering community on reactive systems, are related to our topic but they have not conceptualized *interaction* as such.

mobile systems (Milner, 1993; 1999). To the best of our knowledge, there are three areas where interaction models are framed as such. These areas are some works (i) on concurrency by Milner and his followers (Milner, 1999; 2006), (ii) on Reactive Turing Machines (Andersen, Mørk and Sørensen, 1997; Baeten, Luttik and Tilburg, 2013; Van Leeuwen and Wiedermann, 2001; 2006), and (iii) on interaction as a new non-algorithmic computing paradigm (Goldin, Wegner and Smolka, 2006; Wegner, 1997; Wegner and Goldin, 2003). For each of the three identified models, we:

- present the motivation behind it,
- sum up its account for interaction,
- identify its legacy,
- point out issues regarding the understanding of computers qua that model.

We then want to show how these approaches, which belong to a formal approach, cannot provide an answer to the question of what computers are, and for two reasons. On the one hand, these models of computation have focused their attention on whether interactive models are reducible to models of classical computation—par excellence, the Turing machine. Proving (or not) that an interactive property can be formalized as a computable property in the classical Turing's sense does not answer the question of how an interactive property comes into existence and can be the object of execution. On the other hand, and this is a correlate, these models do not propose a basis for a mechanistic explanation of the very possibility of an interactive computing system. With only formal models of interactive computation, we might run the risk of not offering an adequate conceptualization

of current computers. Therefore, we end up proposing to take the distinction seriously between *models of computation* and *mechanistic computational explanations*, as presented by Miłkowski (2011; 2014).

1. Milner introduces a distinction between interactional and computational behavior

1.1 Motivation

Milner was the one who introduced the concept of interaction in computer science. He summarized his motivations in a famous Turing Award speech (Milner, 1993). Milner was concerned with the logical foundations of computing inherited from Turing. He was preoccupied with the idea that computing practices had evolved since the birth of computing, notably in terms of architecture. He took seriously the possibility that the logical foundations dating back to the thirties may not match the growing challenges of his time and may require additional concepts.

Milner (2006) pointed out that the logical foundations of computing offered by Turing (1937) were previous to the first physical computers and that computer science is grounded in *logic* and *engineering*. On the engineering side, computer science was inherited from von Neumann's pioneering work (Aspray, 1990; Godfrey and Hendry, 1993). Only one thing could happen at once in an early von Neumann's computer. Nevertheless, there was more to computing than von Neumann's architecture (Backus, 1978; Milner, 2006). A growing interest in dealing with concurrency in the sixties and seventies made sequential programming less warranted. Therefore, to Milner, the logical foundations of computing were to evolve. The main flaw of the early logical foundations was the reduction of computing processes

to the concept of an algorithm, which tends to associate computing with mere calculation without taking concurrent activity into account. Because of the evolution of computing engineering practice, Milner questioned whether the logical grounds of computing should evolve as well. Milner's thesis can be put in a nutshell: "this logical foundation has changed a lot since Turing but harks back to him. To be more precise: (i) Computing has grown into informatics—the science of interactive systems; (ii) Thesis: Turing's logical computing machines are matched by a logic of interaction" (Milner, 2006). Consequently, a theory and new language to express concurrent activity were required: "we must find an elementary model which does for interaction what Turing's logical machines do for computation" (Milner, 2006).

The need to define a new computing theory is first displayed through the evolution of computing practice. To sum up, Milner's motivation and focus were the solving of concurrency issues in distributed systems, with the idea that the evolution of computing practices required new formal tools: "Through the 1970s, I became convinced that a theory of concurrency and interaction requires a new conceptual framework, not just a refinement of what we find natural for sequential [algorithmic] computing" (Milner, 1993).

1.2 Account for interaction

Milner introduced the opposition between *interactional* and *computational* behaviour. Introducing the concept of interaction, Milner (1975; 1982; 1983) referred to concurrent message passing between agents. Milner's work coincided with Petri's (1980) new model of concurrent processes, which generally intended to describe concurrency in information systems.[2] To Milner, interaction is more *expressive* than a TM,

[2] Concurrency theory emerged from Dick Karp's early work in the 1960s, grew with (Petri, 1980) and later work on transition systems (Glabbeek and Plotkin, 2004; Nielsen,

but it still describes an *effective* procedure. Milner did not assert *equivalence* between an interactive model and a TM, but he introduced the topic (Milner, 1999) and left it unanswered. Four main differences between old (computational) and new (interactional) computing are made striking by Milner. First, in Milner's words, a Turing Machine prescribes a behaviour to be executed. In contrast, new computing requires the description of an information flow between several system components. Second, old computing is characterized by a *hierarchical design* when current practice involves *heterarchical* phenomena in the computing system. Third, in new computing, the designer cannot predict when agents will be triggered or the overall behaviour of the computing system. Fourth, the user is not merely looking for an end result in new computing practice. There is more than a mathematical function to evaluate, as it used to be in old computing. The user instead interacts with the system, and the look for an end result is replaced by continuing interaction. Having taken stock of the evolution of computing practice on the engineering side, Milner examines its consequences on the logic foundations of computing. The pi-calculus and his work on the equivalence with automata, known as bisimulation, achieved this reflection on interactive processes (Milner, 1993; 1999) with a formalism.

1.3 Legacy

Milner's work on interaction has become a founding block in automata theory and concurrency theory. It installed the notion of a transition system as the prime mathematical model to represent discrete behavior (Arbach et al., 2015; Baldan, Corradini and Montanari, 2001;

Plotkin and Winskel, 1981), and has now developed into a mature theory of reactive systems (Harel and Pnueli, 1985) with diverse network models (for an overview, see Lee and Sangiovanni-Vincentelli, 1998; Lee and Neuendorffer, 2006).

Glabbeek and Plotkin, 2004; Nielsen, Plotkin and Winskel, 1981). It also showed that language equivalence was not the correct notion when comparing automata for interactive systems. Instead, it should be replaced by a notion of behavioral equivalence or bisimilarity (Milner, 1999). The pi-calculus has inspired research to derive a language from it. The *Pict* (Pierce and Turner, 2000) programming language is an example.

Milner's work is foundational and served as a reference for anyone after him, reflecting on the need for a new framework dedicated to new emerging computing practices. Milner insists on an essential reminder that we would like to consider. When modelling, the engineering practice matters and is to be articulated with the logical foundations of the model, should it involve elaborating a new framework. Famously, Wegner and Goldin acknowledge that Milner was the first to introduce the idea that classic models of computation were insufficient. They argue that Milner did not state clearly whether computation in concurrent communicating systems (CCS) and the pi-calculus were reducible to Turing machines and algorithms (Wegner and Goldin, 2003). If one goes and looks at Milner's Turing Award Speech, it seems true that classical computation translates into an interactive calculus. However, it is not stated whether any formula in the pi-calculus can be expressed in a classical calculus like the lambda-calculus.

1.4 Issues for an account of current computers

Given the account of current computers that we are looking for, we see two limits in the lessons drawn from Milner. First, we are looking for an explanation of the interactive computing phenomena at stake in a computer. Therefore, the relation between layers of abstraction, from

the computational to the physical, is crucial. However, to Milner, the physical layer of the machine is not of much interest, and the calculus of CCS needs to abstract away from the physical. As Milner puts it, informatics is about virtual symbols: "physical systems tend to have permanent physical links; they have fixed structure. But most systems in the informatic world are not physical; their links may be virtual or symbolic" (Milner, 1999). From our perspective, abstracting away from the physical world comes at some cost for an explanation. A complete explanation of a computing system can hardly be provided in details within a single understandable abstraction, since a computing sytems is extremely multi-layered (Lee, 2020; Nisan and Schocken, 2005). Therefore, an explanation of a computing system is necessarily a trade-off between understandability and overwhelming details. As we will flesh out in the last section 4 by referring to Miłkowski's work (Miłkowski, 2011; 2016), a good computational explanation must link the formal story and the blueprint of the computing mechanism. Such articulation is not told in a formal theory of concurrent processes. Second, this first story of interactive systems restricts them to concurrent systems, which is only one dimension of interest when describing what current computers do. There are *at least* two core dimensions left aside: what makes possible timing instructions and the connection between physical processes inside and outside the computing system.

2. Reactive TMs: extending the original model

2.1 Motivation

More recently, a literature domain focused on a "Reactive Turing machine" has emerged (Andersen, Mørk and Sørensen, 1997; Baeten, Luttik and Tilburg, 2012; 2013; Luttik and Yang, 2016; Van Leeuwen

and Wiedermann, 2006). It reminds us that the purpose of Turing's *a-machine* model was to propose a formal account of what is *computable by effective means* (algorithmically computable). This formalization was achieved before the realization of the first digital computers. In a way reminiscent of Milner, the question is whether the TM model still fits computing practices decades later. The strategy chosen is to see whether *extensions* of the original TM are sufficient to describe new computing practices and whether the obtained model is still equivalent to a TM. The strategy founds its frame within computability theory and reflects on its scope. This literature domain that proposes extensions of the Turing machine to account for interactive computing systems may be traced back to seminal works on a "Universal reactive machine" (Andersen, Mørk and Sørensen, 1997). In that respect, although pointing at the specificity of interactional behaviour, the main framework still relates to Turing's. Baeten, Luttik, and van Tilburg (Baeten, Luttik and Tilburg, 2013) are looking for a model of interactive computation, extending the classical TM with a process-theoretical notion of interaction related to Milner's previous work. The strategy involves questioning the relationship between such extensions and the Church-Turing thesis. As a reminder, the Church-Turing thesis states that a computable function by effective means is computable by a Turing machine. The community interested in Reactive Turing machines asks the following question: can the Church-Turing thesis also be extended? Van Leeuwen and Wiedermann (2001) focus on the possible extension of the Church-Turing thesis to account for interactive computing: "We will motivate the need for a reconsideration of the classical Turing machine paradigm and formulate an extension of the Church-Turing thesis" (Van Leeuwen and Wiedermann, 2001).

What is at stake is whether the Church-Turing thesis holds given warranted new models of computation: "Is the Church-Turing thesis

as we know it still applicable to the novel ways in which computers are now used in modern information technology? Will it hold for the emerging computing systems of the future?" (Van Leeuwen and Wiedermann, 2001). The Church-Turing thesis originally did not entail a claim about computing in general (what computers do and will do) but only about *effective computation*. Therefore, it does not follow that we should ask the Church-Turing thesis for answers on what computing is. Replacing a question about computing with a question about computation is the mark of a specific formal perspective within the frame of computability theory. Understanding *computing* and its evolution from a formal perspective consist of questioning *what can be computed* and seeing if there is another notion of computation than effective computation *in the sense of Church-Turing*.

2.2 Account for interaction

The starting point in the Reactive TM community is a standard current computer designed as a distributed system interacting with an environmental agent: a *site machine*. Starting from this model, the reflection on interaction aims at showing the equivalence between this site machine and a Turing machine augmented by some functions. The conclusion is that a site machine computer computes effectively and yet requires a TM with new functions, thus requiring an extension of Church-Turing's thesis. There are effectively computable functions that TMs, in a strict sense, cannot compute. One crucial dimension that the community wants to account for is particularly relevant to us: "In order to mimic site machines, a Turing machine must have a mechanism that will enable it to model the change of hardware or software by an operating agent" (Van Leeuwen and Wiedermann, 2001). To make interaction with an external agent possible, the model

needs to integrate a way of entering new, external, and possibly non-computable information into the machine. This is precisely what oracles do. The authors prefer a more general notion: an *advice function*. The model of a Reactive TM (also called a TM with advice) is considered expressive and definitionally equivalent to an Oracle Turing Machine.

Van Leeuwen and Wiedermann identify three key elements that should be integrated all together within the frame of algorithmic computability: "non-uniformity of programs", "interaction of machines", and "infinity of operation". By the non-uniformity of programs, the authors refer to the fact that current programs on a personal computer are no longer fixed but evolve, are upgraded, and their data remain in memory even when the machine is not running. By interaction, they intend to contrast a TM, where all input data are present before the start of the computing procedure, with a modern computer, where continuous streaming of data via input ports is going on. The third mentioned characteristic, the infinity of operation, refers to distributed and mobile communication systems. These systems are to be seen as dynamic networks of many entities sending and receiving signals in unpredictable ways that are to be synchronized. To accommodate the original TM model, Leeuwen and Wiedermann propose to define "Interactive Turing machines with advice." Integrating an "advice" function amounts to entering new, external, and non-computable information into the machine, which requires using oracles (Balcázar, Díaz and Gabarró, 1995; Rogers, 1987). This way, a TM with advice resembles site machines and I/O automata in being equipped with input and output ports. To the authors, formal tools to support interaction and infinite computations are already available. As for interaction, they refer to already well-known and developed literature on the theory of concurrent processes, the programming of parallel processes,

communication protocols, and distributed algorithms. As for infinite computations, Leeuwen and Wiedermann understand them from the language-theoretic viewpoint in the theory of omega-automata (Staiger, 1997; Thomas, 1990).

2.3 Legacy

This approach to extending the Turing machine and the Church-Turing thesis is at the junction between Milner's work and Wegner's (presented in the coming section 3). It makes the junction in that it begs the question of a new paradigm. Milner had not formulated his theory of interaction in such radical terms, but Wegner goes further. The Reactive Turing Machine community asks whether the mentioned required extensions lead to a new computing paradigm: "The experience with present-day computing confronts us with phenomena that are not captured in the scenario of classical Turing machines" (Van Leeuwen and Wiedermann, 2001). The computations carried out on Turing machines with advice are said to be "more powerful" than classic computations on a-machines. The authors insist that this claim does not go against the Church-Turing thesis. To Leeuwen and Wiedermann, like other physical systems (Pour-El, 1999), TMs with advice or oracle Turing machines do not fit the concept of a finite algorithm that can be computed by means of a TM. The conclusion pushes towards a paradigm shift:

> What makes them non-fitting under the traditional notion of algorithms is their potentially endless evolution in time. This includes both interaction and non-uniformity aspects. This gives them the necessary infinite non-uniform dimension that boosts their computational power beyond that of standard Turing machines (Van Leeuwen and Wiedermann, 2001).

The authors ensure that such a paradigm shift does not put into question the original Church-Turing thesis because their proposal for interactive computation does not involve solving undecidable problems (Van Leeuwen and Wiedermann, 2001) using effective computation. The work seems to have served as a pivotal point in structuring the debate on a model of interactive computation around its implications for the Church-Turing thesis. This is evidenced by the objections formulated against Wegner's work which pushes further the concept of interaction and the need for a new paradigm: a proposal of this kind had fallen under objections framed within the theory of computability.

2.4 Issues for an account of current computers

The project is focused on extending the original TM to make it "reactive". The proposed level of abstraction cannot account for the mechanisms that make the proposed extensions possible. We can take a closer look at the type of description presented in this formal framework to account for an interactive scenario:

> The computational scenario of an interactive Turing machine is as follows. The machine starts its computation with empty tapes. It is driven by a standard Turing machine program. At each step, the machine reads the symbols appearing at its input ports. At the same time, it writes some symbols to its output ports. Based on the current context, i.e., on the symbols read on the input ports and in the 'window' on its tapes, and on the current state, the machine prints new symbols under its heads, moves its windows by one cell to the left or to the right or leaves them as they are, and enters a new state. Assuming there is a move for every situation (context) encountered by the machine, the machine will operate in this manner forever. Doing so, its memory (i.e., the amount of rewritten tape) can

> grow beyond any limit. At any time $t > 0$, we will also allow the machine to consult its advice, but only for values of at most t (Van Leeuwen and Wiedermann, 2001).

If we look for a mechanistic explanation of computing, we need some elements to be unpacked beyond a formal account to make sense of the quoted scenario above. For example, we need to account for how reading and writing on the ports are possible. It presupposes that the interactive computing system can wait, pause, and react depending on the arrival or absence of new data. What allows such behavior? It presupposes some mechanisms allowing the system either to be interrupted by environmental processes or to check the new incoming values steadily.[3] In other words, given the initial question ("what is an interactive computer?"), some phenomena cannot be accounted for within the frame of an extended Turing machine. The way oracles work remains at a level of abstraction too remote from the minimal causal blueprint we need for our purpose.

3. Going beyond TMs? Wegner's new paradigm

3.1 Motivation

A strong motivation for Wegner's view on interaction is to overcome the Strong Church-Turing thesis (CTT) that he takes to prevent us from fully admitting a new paradigm in computer science. A paper fleshes out in detail clarifications against the CTT:

> The classical view of computing positions computation as a closed-box transformation of inputs (rational numbers or

[3] More on these mechanisms and on the limitations of oracles can be found in (Martin, Magnaudet and Conversy, forthcoming).

> finite strings) to outputs. According to the interactive view of computing, computation is an ongoing interactive process rather than a function-based transformation of an input to an output. Specifically, communication with the outside world happens during the computation, not before or after it. This approach radically changes our understanding of what computation is and how it is modelled. The acceptance of interaction as a new paradigm is hindered by the Strong Church-Turing Thesis (SCT), the widespread belief that Turing Machines (TMs) capture all computation, so models of computation more expressive than TMs are impossible (Goldin and Wegner, 2008).

In other words, the strong CTT stipulates that a TM could solve all computational problems and could compute anything that any computer can compute. Wegner argues that Turing himself would have denied it, referring to Turing's famous paper (Turing, 1937), as he did not only introduce TMs (calling them automatic machines, or a-machines) but did also introduce choice machines (*c-machines*), extending TMs by allowing a human operator to make choices during the computation. Turing did not view *c-machines* as reducible to TMs, suggesting other forms of computation might exist. Goldin and Wegner also like to remind us that the CTT applies only to the computation of functions rather than to all computations:

> Function-based computation transforms a finite input into a finite output in a finite amount of time, in a closed-box fashion. By contrast, the general notion of computation includes arbitrary procedures and processes—which may be open, nonterminating, and involving multiple inputs interleaved with outputs (Goldin and Wegner, 2008).

For the sake of clarity, Goldin and Wegner propose to formulate the assumptions of the CTT in their proper formulation free of extrapolation (Goldin and Wegner, 2008) explicitly:

- i. "All algorithmic problems are function-based."
- ii. "All function-based problems can be described by an algorithm."
- iii. "Algorithms are what early computers used to do."
- iv. "TMs serve as a general model for early computers."
- v. "TMs can simulate any algorithmic computing device."
- vi. "TMs cannot compute all problems, nor can they do everything that real computers can do."

One reason the strong CTT is "impossible" (Eberbach, Goldin and Wegner, 2004) is that no computable function would determine, given some finite amount of a priori information, all the real-world factors that are necessary to ensure the safe arrival of a car at its destination. An assertion to the contrary would endow TMs with the power to predict the future. Therefore, Wegner introduced *interaction* as a new paradigm, based on an empiricist approach (Wegner, 1995), to broaden algorithmic problem-solving. The reason is that Wegner and his followers take computing machines to be about physical processes, chaotic in nature (Siegelmann, 1995), requiring demanding precision to be controlled (Hartmanis, 1994). Superposed layers of abstractions allow us to describe and control those physical and chaotic computing machines. The challenge is then to bridge the gap between all those layers of abstraction, starting with the lowest physical level. A typical problem we want to solve with computers but not computable in the classic sense would be, e.g., the problem of driving home:

> the problem of driving home from work is computable—by a control mechanism, as in a robotic car, that continuously

> receives video input of the road and actuates the wheel and brakes accordingly. This computation, just as that of operating systems, is interactive, where input and output happen during the computation, not before or after it (Goldin and Wegner, 2008).

Goldin and Wegner argue that such a notion of computation does find its counterpart neither in the theory of computation nor in the concurrency theory. The motivation that goes hand in hand with this discussion against the strong CTT is a reflection on algorithms and the scope of algorithmic problem-solving. Knuth has given a classic definition for algorithms: "An algorithm has zero or more inputs, i.e., quantities which are given to it initially before the algorithm begins" (Knuth, 1968). Following a recipe (Knuth, 1968), for example, does not actually involve algorithmic problem-solving. To know how to mix the ingredients properly, one needs to adapt to dynamic variables and feedback, such as humidity conditions and the progressive evolution of the texture of the paste that are not pre-given values before execution. To Wegner, that kind of feedback does not belong to the function-based mathematical worldview. The problem of driving home from work, like baking following a recipe, is also among those problems that Knuth meant to exclude from his definition.

3.2 Account for interaction

This leads us to Wegner's account for interaction:

> Computational problem solving requires open testing of assertions about engineering problems beyond closed-box mathematical function evaluation. Therefore, we have proposed interactive computing as an empiricist model that expands computational problem solving from algorithmic TM models

> and functional input-output to broader concepts of interleaved dynamic streams and observable interaction with the environment (Wegner and Goldin, 2006).

In Wegner's perspective, interactions are more powerful than TMs with finite initial inputs. TMs with oracles and unbounded (dynamically extensible) input streams model more accurately interactive systems than traditional Turing machines. Interactive systems react dynamically to external events. They are also related to the passage of external time. By delaying the binding time of inputs so that they can occur during the computation (rather than only at the beginning) and modelling reactive processes (Manna and Pnueli, 1992) by infinite computations (Thomas, 1990), the modelled entities are extended from algorithms to persistent objects and concurrent processes (Milner, 1999).

Wegner wonders if Milner himself avoided questioning whether the computation in CCS and the pi-calculus went beyond Turing machines and algorithms (Wegner and Goldin, 2003). The question could remain whether Wegner takes interaction as a super-calculus/super-algorithm or as a radical shift from TMs. In other words, to what extent is "interaction more powerful than algorithm" (Wegner, 1997)? In fact, Wegner's claim is sharp. In contrast with Milner, Wegner's focus is not on concurrency between computing processes. Instead, he focuses on the complexity of the triggering of external events outside the machine: "Interactive systems are grounded in an external reality both more demanding and richer in behaviour than the rule-based world of non-interactive algorithm" (Wegner, 1997). He strikes the difference between closed and opened systems, the latter being impossibly wholly described. This impossibility makes interactive systems mathematically problematic: they lack completeness.

> The comfortable completeness and predictability of algorithms is inherently inadequate in modelling interactive computing tasks and physical systems. The sacrifice of completeness is frightening to theorists who work with formal models like Turing machines [. . .]. But incomplete behaviour is comfortably familiar to physicists and empirical model builders. Incompleteness is the essential ingredient distinguishing interactive from algorithmic models of computing and empirical from rationalist models of the physical world (Wegner, 1997).

From this, Wegner concludes that computing systems should not be thought of as algorithms but as *interfaces*, *views*, and *modes of use*, definable as behaviours to be specified. Consequently, an ontological question is also at stake: in what terms should the external world be modelled: as atomic objects and events? As processes and flow? Formally, Wegner's account of interaction has led to the development of Persistent Turing machines (PTMs), a model of sequential computation, and the result that multi-stream interaction machines (MIMs) are more expressive than sequential interaction machines (SIMs) (Goldin, 2000; Goldin, Smolka et al., 2004). Wegner and Goldin trace back the idea that interaction is not expressible by or reducible to algorithms at the closing conference on the 5th-Generation Computer Project in the context of logic programming. Reactiveness of logic programs, realized by the commitment to a course of action, was shown to be incompatible with logical completeness (Wegner and Goldin, 1999).

3.3 Legacy

Wegner's work has been criticized, the main objection being that interaction machines can be proved equivalent to TMs. The objections are focused on the defence of the Church-Turing thesis (Cockshott and Michaelson, 2007; Prasse and Rittgen, 1998), and assume that

introducing an interactive computing paradigm denies the results of Church and Turing's work. But this assumption cannot be taken for granted: no one denies that TMs and lambda calculus account for *effective computation*. Both formalisms define the intuitive notion of an *algorithm*. The Church-Turing thesis will only be shaken once someone presents an alternative formal account of an effective procedure. Due to semantic ambiguities, some have interpreted Wegner's work as challenging the Church-Turing thesis. First, Wegner characterizes interaction as *more powerful* than algorithms and TMs. What "powerfulness" precisely refers to is unclear. We will say more about this in the next section (section 4).

Second, there seems to be another semantic ambiguity or alleged identity between "computing" and "computation": "Wegner (and Eberbach) say that it is impossible to describe all computations by algorithms. Thus, they do not accept the classic equation of algorithm and effective computation" (Cockshott and Michaelson, 2007). In the former quoted sentence, a core assumption uses interchangeably "computation" and "computing". But Wegner means that it is impossible to describe everything in computing by algorithms. By "computing", he is referring to what computers do broadly, not to Turing computation in a narrow sense. Therefore, the conclusion made in the quoted sentence does not follow: the identity between an effective computation and an algorithm is not put into question by Wegner.

3.4 Issues for an account of current computers

We are interested in the way Wegner broadens the notion of *interaction*. It is not strictly referring to communicating processes within a computing machine. Possible complex interactions with the environment and the dynamic between inputs and outputs during execution

are considered. However, although debunking the focus of the CTT by stating that interaction is more *powerful* and *expressive* than algorithms, Wegner's work is enclosed in a field of discussion framed by the theory of computability. Furthermore, we still need a way of describing the very mechanisms we are interested in to be provided with a mechanistic account of current computers. This is no surprise since Wegner's work aims primarily to reflect on the theoretical limits of classic mathematical tools, e.g., on notions like *completeness*.

4. Why the interactive models identified do not provide us with an answer

We have reviewed the conceptualization of interactive systems in theoretical computer science. We want to defend that these approaches cannot answer the epistemic question asked by philosophers about what current computers are. There are two reasons for this. First, as we have seen, these conceptualizations focus on whether a formal model for interaction is irreducible to a Turing machine and, if so, whether this is a threat to the Church-Turing thesis. This deprives us of a level of description to explain the mechanisms that allow a computing system to be interactive. We propose to detail here in section 4 the problems posed by the debate on reducibility. We end the section by mentioning a distinction currently offered in the literature that highlights the limits of a formal approach. It is a distinction, mostly worked by Miłkowski, opposing *mechanistic computational explanation* and *model of computation*.

	Interaction as concurrent communicating systems	Interaction as extended Turing Machines	Interaction as a new paradigm
Motivation	Provides new logical foundations to fit new engineering challenges, especially concurrency	Extends the TM model to account for interactive devices	Debunks the strong Church-Turing thesis Discusses the scope of algorithmic solving Prones the need for a new computing paradigm
Account for interaction	Information flow Heterarchical design No complete prediction about overall behavior No end-result Process calculi	External data needed during computation Non-uniformity of programs Interaction with agents Infinity of operations Interactive machines are TMs with advice	Computers have rich interaction with the environment during computing execution, but this processing is not merely algorithmic
Uses and criticisms	First conceptualization of interaction Legacy for automata theory	Inspires the need for a new paradigm Puts at the forefront the Church-Turing thesis	Controversy about the powerfulness of the TM
Issues for an account of interactive computing	Definition of interaction restricted to specific properties: concurrency and communication	Formal oracles cannot account for the physical possibility of entering new data	Issues about powerfulness and expressiveness constrict the debate in the realm of computability theory

Table 1: Sum-up: an overview of explicit theories of interactive computing systems in theoretical computer science.

4.1 Unclear stance towards interaction and Turing reducibility

The first problem with the focus on Turing reducibility in the accounts for interaction is that the stance is not always clear-cut. Milner's work leaves us with the following question: to what extent are the new "logical foundations" for interaction distinct from the classic framework? Irreducibility is not stated in the speech for the Turing Award. There

is a simple translation of lambda-calculus into pi-calculus, which is faithful to computational behaviour. Thus, pi-calculus supports functional programming at a higher level of explanation. However, it is unclear whether any behaviour expressed in the pi-calculus can be translated into a classic calculus. In a more recent book, *The Space and Motion of communicating agents* (2009), Milner introduces bigraphs as another formalism for interactive systems. Bigraphs are proven to have the same expressiveness as Turing machines. It looks like Milner proposes to revise the principle of Occam's razor and praise the plurality of formalisms, models, and frames of explanation:

> I reject the idea that there can be a unique conceptual model, or one preferred formalism, for all aspects of something as large as concurrent computation, which is in a sense the whole of our subject — containing sequential computing as a well-behaved special area. We need many levels of explanation: many different languages, calculi, and theories for the different specialisms (Milner, 1993).

It looks like interaction is the new "basic notion":

> Now, what are the new particles, parts of speech, or elements which allow one to express interaction? They lie at the same elementary level as the operation of a Turing machine on its tape, but they differ. For much longer than the reign of modern computers, the basic idiom of algorithm has been the asymmetric, hierarchical notion of operator acting on operand. But this does not suffice to express interaction between agents as peers; worse, it locks the mind away from the proper mode of thought (Milner, 2006).

As for the work on extended Turing Machines, does it involve that interaction is something else, something irreducible to TMs? Does interaction amount to a classical model of computation with extended

computational power? The latter claim is possibly controversial by revising the Church-Turing thesis. In the end, it looks like interaction is still understood in reference to the classical framework (our italics): "examples of interactive [...] indicate that the classical Turing machine paradigm should be *revised (extended)* in order to capture the forms of computation that one observes in the systems and networks in modern information technology" (Van Leeuwen and Wiedermann, 2000).

Criticisms against Wegner show that the criterion of powerfulness is ambiguous when evaluating a model for a computing system. Does powerfulness refer to computational power, involving that an interactive model can express uncomputable functions in Turing's sense? Or does it refer to the expression of more phenomena? Such ambiguity could support some misunderstanding about interaction.

In any case, the literature review on explicit theories of interaction shows that arguments about the powerfulness and equivalence of the interactive and classic models systematically arise.

4.2 Powerfulness and expressiveness: possible ambiguities

Ambiguities around the concepts of powerfulness and expressiveness likely make the debate need clarification. Indeed, there are at least two ways of understanding them. In any case, the powerfulness of a model refers to its expressiveness, which is a semantic property. Expressiveness refers to *what can be expressed* by a given model. If one thinks of a model as a formal language, let us say that expressiveness relates to all the possible sentences one can make in that language.

In a first sense, powerfulness and expressiveness can be understood strictly within computability theory. In that case, the two notions are used when evaluating a mathematical framework supporting the

formalization of semantics. What is called "powerfulness" refers to *computational power*, and expressiveness refers to a formal criterion evaluating which functions can be expressed. *Turing completeness* is then a possible evaluation criterion for expressiveness, for instance.

Let us say that among the things that could be expressed in a model are functions (*set A*) and other things than functions (*set B*). Within each set, some sets include more than others. Within *set A*, the set of hypercomputations is more expressive than the set of computable functions since it includes the uncomputable ones. That is a way to be more expressive: expressing more functions. However, framing expressiveness and powerfulness as possibly only about computable functions would seem odd to engineers and computer scientists familiar with other formalisms than those related to computability theory. Nevertheless, objections about interaction theories frame the debate in reference to computability theory.

In a second sense, one can consider the powerfulness and expressiveness of a model *outside the strictly formal computability framework*. Since a model must represent, according to specific objectives, a phenomenon of reality or, say, a system, we can understand the powerfulness of a model as a good match between the model and what is modeled.

Therefore, in that broader sense, a model is expressive, given some purpose, if and only if it describes all phenomena required for that given purpose. In that case, the value of the model and concerns about its expressiveness depend on stated goals. From an engineering perspective, for example, a model is valuable to the extent that it allows engineers to think of future systems design easily. In this case, the value of the model could be evaluated, e.g., in terms of usability (effectiveness, efficiency, and satisfaction (ISO, 2018)). From a scientific perspective, the aim is to make good predictions about a system. The

two perspectives are rarely used in isolation since good engineering design requires some science, and good science often relies today on some engineering (Lee, 2017). From the perspective of the philosophy of science and given scientific explanation standards, a good model for a phenomenon rightly describes the mechanisms at stake (Glennan, 2002; Machamer, Darden and Craver, 2000; Miłkowski, 2016). Of course, other possible values for models, from other perspectives, could be found.

To go back to Wegner (Goldin, 2000; Wegner, 1995; 1997; 1998), we argue that this distinction between a narrow and broad sense of expressiveness clarifies criticisms made against him.

In a broad sense, one can interpret Wegner's new paradigm as follows: Wegner considers his interactive model more expressive than a TM by having his model describe *other things than Turing computations*. Wegner's model could then describe more phenomena than a TM. It would not go against the Church-Turing thesis, which remains valid to account for algorithmic problem-solving through effective procedures.

But in a narrow sense of expressiveness, one can interpret (wrongly, we think) the possibility of a new paradigm as follows. Wegner and the tenants of Reactive Turing Machines could think of their interactive model as more expressive than a TM, allowing their model to execute *more functions*, even some of them being uncomputable functions in the sense of the Church-Turing thesis, solving the halting problem. In that case, the claim would indeed be controversial. The bold claim would be the following: a TM is not only providing an account for algorithmic problem-solving through effective procedures but it could also be extended to account for other non-algorithmic processes, solving the uncomputable. Interaction would be some super-calculus, extending the calculative power of

the original TM to account for interaction. It would be satisfactorily modeled with a TM, only given more calculation power. It would go down the track of Accelerating Machines or Super-Turing Machines, able to calculate more than Turing's computable functions (Copeland, 2002; Copeland and Shagrir, 2011; MacLennan, 2009).

We argue that a theory of interaction does not need to embrace the hypercomputation view. Part of an interaction model could be reduced to the classical TM, but some extra elements needed to express interaction cannot be reduced to an a-machine. That does not mean interactive models have super computational power to solve undecidable problems. It simply means interactive systems do things that a TM cannot do. It is possible to admit they do other things *without implying they compute uncomputable functions*.

4.3 What formal models of computation cannot do: providing a mechanistic explanation of computing

So, do we have a good theory about interactive computers? Do we understand what they are? A natural and common way to go is to reduce the question of what interactive computers are to what interactive computation is. Initially, the first models of computation emerged through computability theory. They served as answers to an abstract mathematical problem, namely the formalization of the intuitive notion of an algorithm. They had nothing to say about computers, as computers did not even exist at the time. Since the computability era, models of computation like the Turing Machines have been exported outside their original scope to serve as a basis for theoretical computer science. Some models of computation (Turing Machines) have even helped to reflect on computers. It is no surprise since computers were thought to be precisely the kind of machines that implement computations. Models of computation have then evolved, accounting for new

desired properties to be integrated within the classical framework. In computer science, what makes a model of computation valuable is related to the formal properties it expresses. Once those formal properties are at hand, they allow further procedures to be acted upon them, especially system verification and certification. In the end, models of computation serve as tools used to support and verify a system's design. These models belong to a particular abstraction level: they do not intend to model the system as a whole and the way it works. They focus on verifiable properties, upon which proofs that guarantee the outputs of the system are built. Verifying formal properties is different from investigating why the system behaves the way it does. They are two different tasks. The former task (verification) belongs to applied mathematics. It describes abstract computations through formal models by focusing on specific properties. The latter (understanding computing behaviour) is the question the philosopher begs when asking what a computer is. It requires something else than task-oriented formalizations of properties abstracted away from any physical mechanism. Philosophers of computing need to make sense of the overall behaviour, which requires combining other levels of abstraction. The reason is that an account of computing behaviour calls upon the description of how computation can be carried out: in other words, it requires the description of execution on some computer architecture. Computations and their models belong to a level of abstraction independent from implementation detail. Computations, as already coined, are "medium-independent" (Klein, 2020). On the contrary, to have a model of some execution belongs to a lower level of abstraction, where minimal references to the devices that allow the execution are made. There is no need to dig into fine-grained implementation details to make sense of computing behaviour in mechanistic terms.

The formal debate on model equivalence and powerfulness leaves us needing more building blocks to figure out an explanation for interactive computing: what makes it possible, and what mechanisms support it? A helpful distinction here capturing why we lack the right tools is a recent distinction in the literature between models of computation (formal) and mechanistic explanations of computing. It deserves attention in the context of understanding interactive computing. Questioning model equivalence belongs to formal mathematics; it does not aim at providing a mechanistic account of the computing phenomena. Interactive models of computation propose an upper layer of abstractions to formalize specific properties but do not hint at how interactive computation is carried out. We suggest we need to adopt a different explanatory focus, departing from the perspective adopted by models of computation and understanding *how interactive computation can be executed*.

Such lessons have just started to be drawn. They have motivated, for example, distinctions between computational models and computational explanations (Klein, 2020) or between models of computation and computational mechanisms (Miłkowski, 2014). The lesson drawn is that formal models of computing systems do not provide us with the appropriate and complete level of description to build an explanation, which is expected to identify the relevant mechanisms at stake. More precisely, an explanation for computing phenomena requires bridging a high-level description of a computation and its blueprint (Miłkowski, 2011; 2016). The approach is based on the standard of mechanistic explanation in science, coupled with the idea that a computational process is intrinsically mechanistic:

> Computational explanations, according to the mechanistic account are constitutive mechanistic explanations: they explain how a mechanism's computational capacity is generated by

> the orchestrated operation of its component parts. To say that a mechanism implements a computation is to claim that the causal organization of the mechanism is such that the input and output information streams are causally linked and that this link, along with the specific structure of information processing, is completely described (Miłkowski, 2014).

If one is looking for a mechanistic explanation of a computing process, Miłkowski argues that a model of computation may be insufficient. The reason something is missing is that a model of computation is not strongly equivalent to a mechanism:

> There are two ways in which computational models may correspond to mechanisms: first, they may be *weakly equivalent* to the explanandum phenomenon, in that they only describe the input and output information, or *strongly equivalent*, when they also correspond to the process that generates the output information (Miłkowski, 2016).

The difference between strong and weak equivalence captures a difference in causal completeness. The formal models of computation are on the side of models that are weakly equivalent to a mechanism: "formal models cannot function as complete causal models of computers. For example, to repair an old broken laptop, it is not enough to know that it was (idealizing somewhat) formally equivalent to a universal Turing machine." (Miłkowski, 2016). An example helps to flesh out the need for such distinction and turns again to the Turing machine:

> Turing machines were not invented to be implemented physically at all, but some people still build them for fun. [...] Imagine a physical instantiation of a trivial logical negation Turing machine, built of, say, steel and rubber and printing

> symbols on paper tape. Its alphabet of symbols consists of "F" and "T." If the machine finds "T" on its tape, it rewrites it to "F" and halts; if it finds "F," it rewrites it to "T" and halts. Let us suppose that the machine's head is so old and worn out that it tears the paper tape during the readout. As a result, no symbol will appear. [. . .] Only when we describe the Turing machine literally, as a causal system that has a particular causal blueprint (engineering specifications of how it is built), can we causally predict such a breakdown. [. . .] Why are breakdowns and malfunctions so important? They help us discover the causal complexity of the system. [. . .] an abstract model of computation will not predict all the possible outcomes of the breakdown, as it abstracts away from a number of the system's causal characteristics. So it will not tell us what is going to happen with the head; it will only say that the computation will no longer be correct (Miłkowski, 2011).

Thus, Miłkowski invites us to consider a new project in the philosophy of computing: "it is necessary to acknowledge the causal structure of physical computers that is not accommodated by the models used in computability theory" (Miłkowski, 2011). To the best of our knowledge, such a project to account for interactive computing has still not been carried out to flesh out the mechanisms at stake. If philosophers of computing were to proceed in that direction, two criteria for a good explanation of a computer proposed by Miłkowski could offer some guidance. First, such an explanation should be complete, in the sense of a complete causal model where causally relevant parts and operations are specified (Miłkowski, 2014). Second, a good explanation for computing should explain the competence of the system: "By providing the instantiation blueprint of the system, we explain the physical exercise of its capacity, or competence, abstractly specified in the formal model" (Miłkowski, 2014). For example, it would be necessary to be able to explain in mechanistic terms what the behaviour

of an oracle corresponds to. This would be equivalent to explaining which mechanisms allow data arrival, launching, interrupting, or pausing machine processes.

Conclusion

We started from the need to update a question in the philosophy of computing: *what is a computer*? Today's computers are highly interactive, so the question can be rephrased more precisely: *what is an interactive computing system*? It is common to understand computers in terms of existing models of computation, hypothesizing that a computer is primarily a machine that carries out computation. Therefore, the working hypothesis has traditionally answered the initial question by asking what computation is. As already noted, this shift should not be taken for granted. There are, however, and the length of the paper does not allow it, historical and epistemological reasons for this shift that have been described and discussed (Daylight, 2014; Haigh and Priestley, 2020; Mol, 2018). We have chosen in this paper to ask ourselves if the shift is relevant in the case of interactive computing: do we understand what an interactive computer is by questioning the formal models proposed in theoretical computer science for interaction? Our literature review shows that there are better paths. There are two reasons for this. First, the conceptualization of interactive systems in theoretical computer science has focused on their comparison with the Turing machine (and sometimes other classical models), putting forward formal questions about powerfulness and equivalence of models that do not clarify the singularity of interactive systems from an epistemic point of view (rather than formal). There is an inherent difficulty in looking for an explanation of a computing

phenomenon in a formal model: it needs more bricks to describe the mechanisms at stake, at a level of abstraction operating the junction between a high-level formalism and the blueprint. This work does not lead us to an aporia but to a research program in the philosophy of computing: we must produce the right level of explanation for interactive computing.[4] This implies an identification of the mechanisms at play that make possible the interaction between processes within the computing machine (whether there are to be thought of as physical or computational processes, or a mix of both[5]) and the environment. The components of such a mechanism are to be identified and described at a level of abstraction that allows a satisfactory reference to the implementation.

Bibliography

Andersen, H.R., Mørk, S. and Sørensen, M.U., 1997. A universal reactive machine. In: A. Mazurkiewicz and J. Winkowski, eds. *Concur '97: concurrency theory. concur 1997* [Online]. Vol. 1243, Lecture notes in computer science. Berlin; Heidelberg: Springer, pp.89–103. https://doi.org/10.1007/3-540-63141-0_7.

Arbach, Y., Karcher, D., Peters, K. and Nestmann, U., 2015. Dynamic causality in event structures. In: S. Graf and M. Viswanathan, eds. *Formal techniques for distributed objects, components, and systems* [Online]. Vol. 9039, Lecture notes in computer science, pp.83–97. https://doi.org/10.1007/978-3-319-19195-9_6.

Aspray, W., 1990. *John von neumann and the origins of modern computing*. Cambridge, MA: MIT Press.

[4] More considerations on such a research program are fleshed out in (Martin, Magnaudet and Conversy, forthcoming).

[5] The distinction between computational and physical processes is out of the scope of this paper but more on this can be found e.g., in (Kycia and Niemczynowicz, 2020).

Backus, J., 1978. Can programming be liberated from the von Neumann style? *Communications of the ACM*, 21(8), pp.613–641.

Baeten, J.C., Luttik, B. and Tilburg, P.V., 2012. Turing meets milner. *Concur'12: proceedings of the 23rd international conference on concurrency theory* [Online], Lecture notes in computer science. Berlin; Heidelberg: Springer, pp.1–20. https://doi.org/10.1007/978-3-642-32940-1_1.

Baeten, J.C., Luttik, B. and Tilburg, P.V., 2013. Reactive Turing machines. In: O. Owe, M. Steffen and J. Telle, eds. *Fct 2011: fundamentals of computation theory* [Online], Lecture notes in computer science. Berlin; Heidelberg: Springer, pp.348–359. https://doi.org/10.1016/j.ic.2013.08.010.

Balcázar, J.L., Díaz, J. and Gabarró, J., 1995. *Structural complexity I.* [Online]. Second Edition, *Texts in Theoretical Computer Science. An EATCS Series*. Berlin; Heidelberg: Springer. https://doi.org/10.1007/978-3-642-97062-7.

Baldan, P., Corradini, A. and Montanari, U., 2001. Contextual petri nets, asymmetric event structures, and processes. *Information and Computation* [Online], 171(1), pp.1–49. https://doi.org/10.1006/inco.2001.3060.

Basman, A., Tchernavskij, P., Bates, S. and Beaudouin-Lafon, M., 2018. An anatomy of interaction: co-occurrences and entanglements. *Conference companion of the 2nd international conference on art, science, and engineering of programming* [Online]. Association for Computing Machinery, pp.188–196. https://doi.org/10.1145/3191697.3214328.

Beaudouin-Lafon, M., 2006. Human-computer interaction. In: D. Goldin, S.A. Smolka and P. Wegner, eds. *Interactive computation: the new paradigm* [Online]. Berlin; Heidelberg: Springer, pp.227–254. https://doi.org/10.1007/3-540-34874-3_10.

Bielecki, A., 2019. *Models of neurons and perceptrons: selected problems and challenges* [Online]. Vol. 770, *Studies in Computational Intelligence*. Springer International Publishing. https://doi.org/10.1007/978-3-319-90140-4.

Cockshott, P. and Michaelson, G., 2007. Are there new models of computation? Reply to Wegner and Eberbach. *Computer Journal* [Online], 50(2), pp.232–247. https://doi.org/10.1093/comjnl/bxl062.

Copeland, B.J., 2002. Hypercomputation. *Minds and Machines* [Online], 12, pp.461–502. https://doi.org/10.1023/A:1021105915386.

Copeland, B.J. and Shagrir, O., 2011. Do accelerating Turing machines compute the uncomputable? *Minds and Machines* [Online], 21, pp.221–239. https://doi.org/10.1007/s11023-011-9238-y.

Daylight, E.G., 2014. A Turing tale. *Communications of the ACM* [Online], 57(10), pp.36–38. https://doi.org/10.1145/2629499.

Dearden, A.M. and Harrison, M.D., 1997. Abstract models for HCI. *International Journal of Human Computer Studies* [Online], 46(1), pp.151–177. https://doi.org/10.1006/ijhc.1996.0087.

Dodig-Crnkovic, G., 2011. Significance of models of computation, from Turing model to natural computation. *Minds and Machines* [Online], 21, pp.301–322. https://doi.org/10.1007/s11023-011-9235-1.

Eberbach, E., Goldin, D. and Wegner, P., 2004. Turing's ideas and models of computation. In: C. Teuscher, ed. *Alan turing: life and legacy of a great thinker* [Online]. Berlin; Heidelberg: Springer, pp.159–194. https://doi.org/10.1007/978-3-662-05642-4_7.

Glabbeek, R.V. and Plotkin, G., 2004. Event structures for resolvable conflict. *29th international symposium on mathematical foundations of computer* [Online]. Vol. 3153, Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Berlin; Heidelberg: Springer, pp.550–561. https://doi.org/10.1007/978-3-540-28629-5_42.

Glennan, S., 2002. Rethinking mechanistic explanation. *Philosophy of Science* [Online], 69(S3), pp.342–353. https://doi.org/10.1086/341857.

Godfrey, M.D. and Hendry, D.F., 1993. The computer as von neumann planned it. *IEEE Annals of the History of Computing* [Online], 15(1), pp.11–21. https://doi.org/10.1109/85.194088.

Goldin, D., 2000. Persistent Turing machines as a model of interactive computation. In: K. Schewe and B. Thalheim, eds. *Foundations of Information and Knowledge Systems. FoIKS 2000* [Online]. Vol. 1762, Lecture notes in computer science. Berlin; Heidelberg: Springer, pp.116–135. https://doi.org/10.1007/3-540-46564-2_8.

Goldin, D., Smolka, S.A., Attie, P.C. and Sonderegger, E.L., 2004. Turing machines, transition systems, and interaction. *Information and Computation* [Online], 194(2), pp.101–128. https://doi.org/https://doi.org/10.1016/j.ic.2004.07.002.

Goldin, D. and Wegner, P., 2008. The interactive nature of computing: Refuting the strong Church-Turing thesis. *Minds and Machines* [Online], 18, pp.17–38. https://doi.org/10.1007/s11023-007-9083-1.

Goldin, D., Wegner, P. and Smolka, S.A., 2006. *Interactive computation: the new paradigm* [Online]. Berlin; Heidelberg: Springer. https://doi.org/10.1007/3-540-34874-3.

Haigh, T. and Priestley, M., 2020. Historical reflections von neumann thought turing's universal machine was 'simple and neat.' but that didn't tell him how to design a computer. *Communications of the ACM* [Online], 63(1), pp.26–32. https://doi.org/10.1145/3372920.

Harel, D. and Pnueli, A., 1985. On the development of reactive systems. In: K. Apt, ed. *Logics and models of concurrent systems* [Online]. Vol. 13, Nato asi series. Berlin; Heidelberg: Springer, pp.477–498. https://doi.org/10.1007/978-3-642-82453-1_17.

Hartmanis, J., 1994. Turing award lecture on computational complexity and the nature of computer science. *Communications of the ACM* [Online], 37(10), pp.37–43. https://doi.org/10.1145/194313.214781.

Hornbaek, K. and Oulasvirta, A., 2017. What is interaction? *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* [Online]. New York, NY: Association for Computing Machinery, pp.5040–5052. https://doi.org/10.1145/3025453.3025765.

ISO, 2018. Ergonomics of human-system interaction. part 11: usability: definitions and concepts. *ISO 9241-11:2018*.

Klein, C., 2020. Polychrony and the process view of computation. *Proceedings of the 2018 Biennial Meeting of the Philosophy of Science Association. Part II* [Online]. Vol. 87, 5, pp.1140–1149. https://doi.org/10.1086/710613.

Knuth, D.E., 1968. *The art of computer programming, vol.1: fundamental algorithms*. Boston; etc.: Addison-Wesley.

Kycia, R. and Niemczynowicz, A., 2020. Information and physics. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* [Online], (69), pp.237–252. Available at: <https://zfn.edu.pl/index.php/zfn/article/view/513>.

Lee, E.A., 2017. Fundamental limits of cyber-physical systems modeling. *ACM Transactions on Cyber-Physical Systems* [Online], 1(1), pp.1–26. https://doi.org/10.1145/2912149.

Lee, E.A., 2020. *Plato and the nerd* [Online]. MIT Press. https://doi.org/10.7551/mitpress/11180.001.0001.

Lee, E.A. and Neuendorffer, S., 2006. Concurrent models of computation for embedded software. *System-on-Chip: Next Generation Electronics* [Online]. https://doi.org/10.1049/PBCS018E_ch7.

Lee, E.A. and Sangiovanni-Vincentelli, A., 1998. A framework for comparing models of computation. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* [Online]. https://doi.org/10.1109/43.736561.

Luttik, B. and Yang, F., 2016. On the executability of interactive computation. In: A. Beckmann, L. Bienvenu and N. Jonoska, eds. *Pursuit of the universal. cie 2016* [Online]. Vol. 9709, Lecture notes in computer science. Cham: Springer, pp.312–322. https://doi.org/10.1007/978-3-319-40189-8_32.

Machamer, P., Darden, L. and Craver, C.F., 2000. Thinking about mechanisms. *Philosophy of Science* [Online], 67(1), pp.1–25. https://doi.org/10.1086/392759.

MacLennan, B., 2003. Transcending Turing computability. *Minds and Machines* [Online], 13, pp.3–22. https://doi.org/10.1023/A:1021397712328.

MacLennan, B., 2009. Super-Turing or non-Turing? extending the concept of computation. *International Journal of Unconventional Computing*, 5(3-4), pp.369–387.

Manna, Z. and Pnueli, A., 1992. *The temporal logic of reactive and concurrent systems* [Online]. Springer. https://doi.org/10.1007/978-1-4612-0931-7.

Martin, A., Magnaudet, M. and Conversy, S., forthcoming. Computers as interactive machines: Can we build an explanatory abstraction? *Minds and Machines*.

Milner, R., 1975. Processes: A Mathematical Model of Computing Agents. In: H.E. Rose and J.C. Shepherdson, eds. *Studies in Logic and the Foundations of Mathematics* [Online]. Vol. 80, *Logic Colloquium '73*. Elsevier, pp.157–173. https://doi.org/10.1016/S0049-237X(08)71948-7.

Milner, R., 1982. Four combinators for concurrency. *Proceedings of the annual acm symposium on principles of distributed computing* [Online], Podc '82. New York, NY: Association for Computing Machinery, pp.104–110. https://doi.org/10.1145/800220.806687.

Milner, R., 1983. Calculi for synchrony and asynchrony. *Theoretical Computer Science* [Online]. https://doi.org/10.1016/0304-3975(83)90114-7.

Milner, R., 1993. Elements of interaction: Turing Award Lecture. *Communications of the ACM* [Online], 36(1), pp.78–89. https://doi.org/10.1145/151233.151240.

Milner, R., 1999. *Communicating and mobile systems: the π-calculus*. Cambridge: Cambridge University Press.

Milner, R., 2006. Turing, computing and communication. In: *Interactive Computation: The New Paradigm* [Online]. Ed. by D. Goldin, S.A. Smolka and P. Wegner. Berlin; Heidelberg: Springer, pp.1–8. https://doi.org/10.1007/3-540-34874-3_1.

Milner, R., 2009. *The space and motion of communicating agents*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511626661.

Miłkowski, M., 2011. Beyond formal structure: a mechanistic perspective on computation and implementation. *Journal of Cognitive Science*, 12, pp.359–379.

Miłkowski, M., 2014. Computational mechanisms and models of computation. *Philosophia Scientiae* [Online], 18(3), pp.215–228. https://doi.org/10.4000/philosophiascientiae.1019.

Miłkowski, M., 2016. A mechanistic account of computational explanation in cognitive science and computational neuroscience. In: V.C. Müller, ed. *Computing and Philosophy: Selected Papers from IACAP 2014* [Online]. Vol. 375. Springer International Publishing, pp.191–205. https://doi.org/10.1007/978-3-319-23291-1_13.

Mol, L.D., 2018. Turing machines. *Stanford Encyclopedia* [Online]. Available at: <https://plato.stanford.edu/entries/turing-machine/>.

Myers, B., 1994. Challenges of HCI design and implementation. *Interactions* [Online]. https://doi.org/10.1145/174800.174808.

Nielsen, M., Plotkin, G. and Winskel, G., 1981. Petri nets, event structures and domains, part I. *Theoretical Computer Science* [Online], 13(1), pp.85–108. https://doi.org/10.1016/0304-3975(81)90112-2.

Nisan, N. and Schocken, S., 2005. *The Elements of Computing Systems: Building a Modern Computer from First Principles*. Cambridge: MIT Press.

Petri, C., 1980. Introduction to general net theory. In: W. Brauer, ed. *Net theory and applications* [Online]. Vol. 84, Lecture notes in computer science. Berlin; Heidelberg: Springer. https://doi.org/10.1007/3-540-10001-6_21.

Piccinini, G., 2008. Computers. *Pacific Philosophical Quarterly* [Online], 89(1), pp.32–73. https://doi.org/10.1111/j.1468-0114.2008.00309.x.

Pierce, B.C. and Turner, D.N., 2000. Pict: a programming language based on the pi-calculus. *Proof, Language and Interaction: Essays in Honour of Robin Milner*. Cambridge, MA: MIT Press, pp.455–494.

Pour-El, M.B., 1999. The structure of computability in analysis and physical theory: an extension of Church's thesis. In: E. Griffor, ed. *Handbook of Computability Theory, Studies in Logic and the Foundations of Mathematics* [Online]. Amsterdam: Elsevier, pp.449–471. https://doi.org/10.1016/S0049-237X(99)80029-9.

Prasse, M. and Rittgen, P., 1998. Why Church's thesis still holds. some notes on Peter Wegner's tracts on interaction and computability. *The Computer Journal* [Online], 41, pp.357–362. https://doi.org/10.1093/comjnl/41.6.357.

Rapaport, W.J., 2018. What is a computer? a survey. *Minds and Machines* [Online], 28(3), pp.385–426. https://doi.org/10.1007/s11023-018-9465-6.

Rogers, H., 1987. *Theory of recursive functions and effective computability*. Cambridge, MA: MIT Press. https://doi.org/10.2307/3614588.

Siegelmann, H.T., 1995. Computation beyond the Turing limit. *Science* [Online], 268(5210), pp.545–548. https://doi.org/10.1126/science.268.5210.545.

Smith, B.C., 2002. The foundations of computing. In: M. Scheutz, ed. *Computationlism: new directions*. Cambridge, MA: MIT Press.

Soare, R.I., 2013. Interactive computing and relativized computability. In: *Computability: Turing, Godel, Church, and Beyond* [Online]. Ed. by B.J. Copeland. Cambridge, MA: MIT Press. Chap. 9, pp.214–271. https://doi.org/10.7551/mitpress/8009.003.0010.

Staiger, L., 1997. Omega-languages. In: *Handbook of formal languages* [Online]. Ed. by G. Rozenberg and A. Salomaa. Berlin; Heidelberg: Springer, pp.339–387. https://doi.org/10.1007/978-3-642-59126-6_6.

Thomas, W., 1990. Automata on infinite objects. In: *Formal models and semantics* [Online]. Amsterdam: Elsevier. Chap. 4, pp.133–191. https://doi.org/10.1016/b978-0-444-88074-1.50009-3.

Turing, A., 1937. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society* [Online], s2-42(1), pp.230–265. https://doi.org/10.1112/plms/s2-42.1.230.

Van Leeuwen, J. and Wiedermann, J., 2000. On the power of interactive computing. In: J. van Leeuwen et al., eds. *Theoretical computer science: exploring new frontiers of theoretical informatics* [Online]. Vol. 1872, *TCS 2000. Lecture Notes in Computer Science*. Berlin; Heidelberg: Springer. https://doi.org/10.1007/3-540-44929-9_48.

Van Leeuwen, J. and Wiedermann, J., 2001. Beyond the turing limit: Evolving interactive systems. In: L. Pacholski and P. Ružička, eds. *SOFSEM 2001: Theory and Practice of Informatics* [Online]. Vol. 2234, Lecture notes in computer science. Berlin; Heidelberg: Springer, pp.90–109. https://doi.org/10.1007/3-540-45627-9_8.

Van Leeuwen, J. and Wiedermann, J., 2006. A theory of interactive computation. *Interactive computation: the new paradigm* [Online]. Berlin; Heidelberg: Springer, pp.119–142. https://doi.org/10.1007/3-540-34874-3_6.

Wegner, P., 1995. Interaction as a basis for empirical computer science. *ACM Computing Surveys (CSUR)* [Online], 27(1), pp.45–48. https://doi.org/10.1145/214037.214092.

Wegner, P., 1997. Why interaction is more powerful than algorithms. *Communications of the ACM* [Online], 40(5), pp.80–91. https://doi.org/10.1145/253769.253801.

Wegner, P., 1998. Interactive foundations of computing. *Theoretical Computer Science* [Online], 192(2), pp.315–351. https://doi.org/10.1016/S0304-3975(97)00154-0.

Wegner, P. and Goldin, D., 1999. Interaction as a framework for modeling. In: G. Goos et al., eds. *Conceptual modeling: current issues and future directions* [Online]. Vol. 1565, *Lecture Notes in Computer Science*. Berlin; Heidelberg: Springer, pp.243–257. https://doi.org/10.1007/3-540-48854-5_19.

Wegner, P. and Goldin, D., 2003. Computation beyond Turing machines. *Communications of the ACM* [Online], 46(4), pp.100–102. https://doi.org/10.1145/641205.641235.

Wegner, P. and Goldin, D., 2006. Principles of problem solving. *Communications of the ACM* [Online], 49(7), pp.27–29. https://doi.org/10.1145/1139922.1139942.

Shannon-inspired information in the clinical use of neural signals concerning post-comatose patients

Hyungrae Noh
Sunchon National University, South Korea

Abstract

Post-comatose patients are classified as being in a minimally conscious state when they have executive functions. Because traditional behavioral assessments may not capture signs of executive functions in post-comatose patients, clinicians look to localized brain activities in response to task instructions, such as imagining wiggling toes, to diagnose minimal consciousness. This paper critically assesses the assumption underlying such alternative methods: that brain activities are neural signals conveying information about minimal consciousness. Based on a Shannon-inspired idea of information, I distinguish between informational and engineering aspects of clinical tasks. The informational aspect concerns the conditional probability that, for example, given activity in the motor areas of the brain in response to task instructions, a patient is imagining wiggling toes. The engineering aspect concerns efficient activation of the relevant brain areas in a patient under the task conditions. This distinction shows that the current alternative methods are not informationally problematic, but are structurally "ill-formed." For instance, the toe-imagery task requires the capacity to comprehend syntactically complex sentences, which can be dissociated from minimal consciousness. I propose a misrepresentation task, which tests the capacity to misconceptualize lukewarm water as melting wax, as a supplement to the current alternative meth-

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 121–145

ods. This task is as informationally reliable as these methods, but is structurally "well-formed," as it does not rely methodologically on prerequisites such as language comprehension.

Keywords

post-comatose disorders of consciousness, Shannon's theory of information, minimal consciousness, mental action.

1. Introduction

Minimally conscious post-comatose patients (henceforth, MCP patients) can have limited executive functions that allow for the patients to follow simple instructions or respond to certain stimulations. Signs of residual executive functions provide clinicians with strong evidence of eventual recovery of consciousness (Naccache, 2018; Rohaut, Eliseyev and Claassen, 2019). Thus, delayed recognition of such signs may lead to suboptimal clinical care (van Erp, Aben et al., 2019), distorted prognostic figures for rehabilitation (Ansell, 1993), or ethical problems (Peterson and Bayne, 2018; Noh, 2022). Such signs, however, might not be captured by traditional behavioral assessments (e.g., eye tracking, speaking, etc.) because limited executive functions can be dissociated from the capacity to perform overt behaviors (Teasdale and Jennett, 1974; Andrews et al., 1996). Although some traditional behavioral assessment methods can improve diagnostic accuracy (Schnakers et al., 2009), differentiating MCP patients from patients in a vegetative state remains challenging (van Erp, Lavrijsen et al., 2015).

As an alternative to behavioral assessments, clinicians can use measures such as functional magnetic resonance imaging (fMRI) or electroencephalograms (EEG) to capture localized brain activities as a mark of residual executive functions. Such neural-signal-based assessments (henceforth NSAs) allow clinicians to diagnose minimal consciousness without appealing to overt behavior. For instance, minimal consciousness can be ascribed to a patient if activities in the motor areas of the brain are observed when the patient is instructed to imagine performing a particular motor action (Owen et al., 2006; Cruse et al., 2011; Wang et al., 2019), or if the P300b response is observed when a series of auditory stimulations is given to the patient (Boly et al., 2011; King et al., 2013). An overarching assumption of NSAs is that purported brain activities can be taken as neural signals conveying information about minimal consciousness.

The aim of this paper is to critically assess whether NSAs satisfy this overarching assumption. Drawing on a Shannon-inspired idea of information, I distinguish between informational and engineering aspects of neural signal processing in the relevant studies. Shannon's (1948) theory of information "defines the quantity of information in a signal over the space of possibilities in a given situation where a sender and a receiver are communicating" (Noh, 2018, p.179). The primary concern of this paper is the communication channel between an MCP patient and a clinician (or a test-subject and an experimenter), where signal-vehicles are localized brain activities measured using fMRI or EEG. I apply Shannon's theory to this communication channel in order to estimate the quantity of information about minimal consciousness possibly conveyed by the neural signal. Following Weaver (1953, p.270), I further discuss whether the communication channel is designed to efficiently handle the assigned jobs. Notice that I use the term "Shannon-inspired idea of information" rather than

"Shannon's theory of information" because Shannon did not intend the theory to be applied to this sort of communication channel, and I make certain assumptions in order to apply it here (see footnotes 2 and 3 for details of these assumptions).

I discuss the informational and engineering aspects of clinical tasks in the paper's second and third sections, respectively. Briefly, the informational aspect concerns the conditional probability that, for example, given activity in the motor areas in response to task instructions, the patient is mentally acting. On the other hand, the engineering aspect concerns efficient activation of the relevant brain areas, specifically in relation to task conditions. For example, any patient response to a task with the verbal instruction "Imagine wiggling all of your toes" requires the patient to have the capacity to properly comprehend the instruction as a command to perform a kinesthetic mental-motor action. As I discuss in the following sections, making this distinction between the two aspects shows that NSAs in their current form are structurally "ill-formed." For instance, brain activities in the mental-motor action task would indicate that the patient is minimally conscious, but the task's applicability is limited because the capacity to comprehend syntactically complex sentences (i.e., the verbal instructions of the task) can be dissociated from minimal consciousness. In the fourth section, I propose a misrepresentation task as a supplement to NSAs in their current form. This task tests whether a subject has the capacity to misconceptualize lukewarm water as melting wax. I claim that this task is as informationally reliable as NSAs in their current form and is structurally "well-formed" in that it is not methodologically reliant on prerequisites (e.g., language comprehension), as are NSAs in their current form.

2. The Informational Aspect of Clinical Tasks

NSAs can be categorized as either active or passive (Kondziella et al., 2016). Patients tested under active paradigms are instructed to mentally act, such as imagining playing tennis or walking from room to room in their home (Owen et al., 2006; Monti et al., 2010), imagining squeezing their hands or wiggling their toes (Cruse et al., 2011), mentally counting occurrences of a target (e.g., words like "yes" or "no") in a sequence of sounds (Lulé et al., 2013; Naci and Owen, 2013), or imagining raising their hands (Wang et al., 2019). On the other hand, patients tested under passive paradigms are instructed to pay attention to a series of auditory sequences (Boly et al., 2011; King et al., 2013) or to watch a movie (Naci, Cusack et al., 2014; Laforge et al., 2020).

An essential difference between active and passive paradigms is that while the former requires a subject to either perform a mental-motor action or count occurrences of a target, the latter concerns whether a subject is able to experience stimulation without performing a mental action. Nonetheless, NSAs in general appeal to localized brain activities, which can be captured either by fMRI or by EEG, as a means of diagnosis of minimal consciousness.[1] Specifically, NSAs assume that the observed brain activities are neural signals conveying the message that the subjects under consideration are following instructions or paying attention to stimulations. To analyze the informational aspect is to assess whether the signals do convey the purported message.

Consider Cruse et al.'s (2011) toe-imagery task in order to analyze the informational aspect of active paradigm tasks. They instructed 16 behaviorally nonresponsive post-comatose patients (who were initially

[1] For the sake of simplicity, I am going to ignore methodological differences in the use of fMRI and EEG, which are irrelevant to my argument.

diagnosed with the vegetative state by traditional behavioral methods) and 12 healthy controls to imagine wiggling all of their toes on both feet and relaxing them without making any actual motor movement. Cruse et al. write that a significant degree of activity was observed in a relatively localized area of the medial premotor cortex of 3 patients and 9 controls. We can distinguish between the minimally conscious mental event, M_1, and a set of nonconscious mental events, M_2, in relation to a brain event, namely the activity in the medial premotor cortex, S.[2] To analyze the informational aspect for active paradigms is to compare M_1 and M_2.[3]

Mutual information between the brain event, S, and a set of mental events, M, is:

$$I(S : M) = H(M) - H(M|S).^{4}$$

[2] Shannon's theory conceptualizes mutual information as the relation between events and a subset of these events. In the body of the text, I assume that mental events (M_1 and M_2) are a subset of a brain event (S). More specifically, I assume that functional correlations between cognitive capacities and activities in the corresponding brain areas (e.g., relationships between particular types of motor actions and activities in the corresponding brain areas) lead to mutual information between the mental events and the brain event. Obviously, this assumption is controversial. Nonetheless, in the current paper, I have made an attempt to demonstrate that such an information-theoretic assumption inevitably leads to a clinically important conclusion. See (Li et al., 2022) for a detailed discussion of the quantitative relationships between localized brain activities and information processing capacities. Kycia (2021) provides good clarification of fundamental issues concerning classical and quantum nature of information storing and processing in the brain.

[3] Strictly speaking, M_1 should be a disjunction of the minimally conscious state and the fully conscious state (in Cruse et al., the healthy controls were fully conscious). Because the primary concern of NSAs is the minimally conscious state, however, we can safely ignore the information about the fully conscious state possibly conveyed by the neural signals.

[4] The idea of analyzing the informational aspect of active paradigm tasks originates from Noh (2018, pp.181–185).

H refers to Shannon entropy. To compare M_1 and M_2, however, individual events need to be considered rather than the average. Because the toe-imagery task depends on the assumption that the detection of S particularly reduces the uncertainty of the mental event that subjects are imagining wiggling toes, let's take M_1 as referring to this mental event. The amount of uncertainty of M_1 that S reduces is:

$$\begin{aligned} I(S : M_1) &= I(M_1) - I(M_1|S) \\ &= -\log_2 P(M_1) + \log_2 P(M_1|S) \\ &= \log_2 \frac{P(M_1|S)}{P(M_1)}. \end{aligned}$$

As $\log_2 \frac{P(M_1|S)}{P(M_1)}$ gets more and more positive, S will represent greater accuracy about M_1. If $\log_2 \frac{P(M_1|S)}{P(M_1)}$ is greater than $\log_2 \frac{P(M_2|S)}{P(M_2)}$, then the observation of S provides clinicians with evidence that patients in the toe-imagery task who show activities in the medial premotor cortex are minimally conscious. If the latter is greater than, or equivalent to, the former, then S alone cannot provide that evidence.

To estimate the value of $\log_2 \frac{P(M_1|S)}{P(M_1)}$, we need to understand the background conditions of the toe-imagery task. Notice that the patients in Cruse et al.'s (2011) experiment were initially diagnosed with the vegetative state by traditional behavioral assessments. Given that the misdiagnosis rate of traditional behavioral assessments is 41% (Schnakers et al., 2009), the unconditional probability that Cruse et al.'s patients were conscious, i.e., $P(M_1)$, is lower than the unconditional probability that they were in the vegetative state, i.e., $P(M_2)$. So, $\log_2 \frac{P(M_1|S)}{P(M_1)}$ is greater than $\log_2 \frac{P(M_2|S)}{P(M_2)}$ if $P(M_1|S)$ is greater than, or equivalent to, $P(M_2|S)$.

$P(M_1|S)$ can be accounted for by the reversed strong correlation that holds between activities in motor areas and the corresponding mental-motor action. According to relevant experiments (e.g.,

Pfurtscheller and Neuper, 1997; Ehrsson, Geyer and Naito, 2003) in which healthy subjects were instructed to mentally act (e.g., they were instructed to imagine raising their left arms without making any actual motor movement), activities in the relevant motor areas were observed in the majority of the subjects. Provided that P(activities in the relevant motor areas|mental-motor action) is high, it seems that $P(M_1|S)$ is relatively high as well.

On the other hand, $P(M_2|S)$ concerns a situation in which the medial premotor cortex activity in a mental-motor action task, such as the toe-imagery task, is somehow automated (i.e., a nonconscious response). Provided the strong functional relationship between the body surface and a localized motor area (i.e., the well-established somatotopic map in motor areas), it seems possible to take the medial premotor cortex activity observed in the toe-imagery task as a toe-related nonconscious mental event. In turn, to account for $P(M_2|S)$ would require finding a toe-related nonconscious explanation (that does not involve volition, linguistic understanding of the verbal instruction, etc.) of why the activity occurred.

Nonetheless, no such explanation is available. Consider a semantic priming effect, which refers to cases where automated activities in the somatotopy of the motor and premotor cortex are observed for a few milliseconds when a subject hears an action-word such as "kick" (Pulvermüller, 2005). Now, consider M_2 as a toe-related nonconscious mental event grounded by a semantic priming effect. $P(M_2|S)$ in this sense, however, is very low for two reasons. First, the semantic priming effect has been observed when a single word is given, but not when a whole sentence, such as "Imagine wiggling all of your toes," is given (Raposo et al., 2009). Second, the effect

lasts for a very short time, but the activity in Cruse et al.'s experiment persisted for more than a few seconds. Because there is no alternative explanation, $\log_2 \frac{P(M_1|S)}{P(M_1)}$ is greater than $\log_2 \frac{P(M_2|S)}{P(M_2)}$.

Following the same line of reasoning to analyze the informational aspect of passive paradigm tasks, let us consider King et al.'s (2013) and Bekinschtein et al.'s (2009) oddball tasks. The oddball task is designed to evaluate cerebral responses to violations of temporal regularities that are global across several seconds. Suppose that a stream of repeated auditory events is given to a healthy subject, with two types of novel events (i.e., oddballs) embedded in the stream. The two types are local oddballs, which consist of a change in pitch within a five-sound sequence (e.g., AAAAB), and global oddballs, which consist of a change in an auditory sequence in a fixed global context (e.g., AAAAB AAAAB AAAAB AAAAB AAAAA). Local oddballs typically lead to so-called frontal mismatch negativity, which is an automated (i.e., nonconscious) brain activity. In other words, frontal mismatch negativity is observed independently of whether a subject pays attention to the given auditory stream or not. On the other hand, global oddballs generate the so-called P300b response, which can be detected only when a subject is consciously paying attention to the given auditory stream. In King et al.'s (2013) experiment, a global effect was found in 14% of 70 vegetative state patients and 31% of 65 minimally conscious state patients (and 52% of 23 conscious controls with brain injuries). In Bekinschtein et al.'s (2009) experiment, a global effect was found in none of 3 vegetative state patients and 3 of 4 minimally conscious state patients (and all of 11 healthy controls). But there was no significant difference between the patient group and the control group regarding the local effect in both experiments.

The oddball task takes the P300b response as a neural signal conveying the message that the subject is in a mental state of counting

the number of global deviant trials by using working memory. To take the neural signal as conveying the purported message is to assume a strong correlation between the P300b response and the relevant executive functions, like working memory.

Recall that in Cruse et al.'s (2011) toe-imagery task, a strong correlation between activities in motor areas and the corresponding mental-motor action holds because $P(M_1|S)$ is relatively greater than $P(M_2|S)$.[5] So, we should compare P(the subject is minimally conscious|the P300b response) and P(the subject is nonconscious|the P300b response). According to the relevant studies, the P300b response requires working memory. Specifically, the P300b response is accounted for by the following hypotheses: Noticing global oddballs requires working memory and predictive coding (Garrido, Friston et al., 2008; Garrido, Kilner et al., 2009); the P300b response is a neural signature of postperceptual processing, such as working memory (Cohen et al., 2020); and the P300b response is a neural signature of working memory–guided categorization processes (Rac-Lubashevsky and Kessler, 2019). If there is no alternative explanation of the P300b response, and if working memory is a sign of minimal consciousness (Ansell, 1993; Bekinschtein et al., 2009; King et al., 2013), then P(the subject is minimally conscious|the P300b response) is greater than P(the subject is nonconscious|the P300b response).

A false positive occurs when evidence of an effect is measured, yet the target phenomenon is absent from the test conditions (Peterson, Cruse et al., 2015, p.591), such as a case in which the observed brain activity does not carry the information about minimal consciousness.

[5] Notice that $P(M_2)$ is always greater than $P(M_1)$ in both active and passive paradigms because every MCP patient in the experiments was initially diagnosed with the vegetative state by traditional behavioral assessments. Specifically, because the misdiagnosis rate of traditional behavior assessments is 41%, $P(M_2)$ is roughly 60%.

The probability that NSAs generate false positives is low because NSAs in general involve a relatively strong correlation between brain activities and minimal consciousness. Specifically, given that there is no plausible alternative explanation of why activities in motor areas or P300b responses occurred in the relevant experiments, the best explanation is the one that appeals to minimal consciousness.

Di et al.'s (2008) cohort study provides an additional reason to take the patients who passed the relevant tests as being in a state of minimal consciousness. They found that post-comatose patients who were re-diagnosed with the minimally conscious state by NSAs had recovered sophisticated cognitive functions or consciousness. This finding indicates that activities in motor areas or P300b responses are indeed prognostic signs of recovery. So, the low possibility of false positives can be ignored for the sake of possible recovery. Consequently, activities in motor areas or P300b responses in the relevant experiments can be taken as neural signals conveying the message that the post-comatose patients under consideration are minimally conscious, and therefore, they were initially misdiagnosed with the vegetative state by traditional behavioral assessments.

3. The Engineering Aspect of Clinical Tasks

The informational and engineering aspects of a communication system are connected to the extent that the system is able to handle any message from a set of possible messages produced by a sender (Weaver, 1953, p.270). Consider a Morse-code-based telegraph system. The informational aspect of the system concerns conventions pertaining to the use of the Morse codes in accordance with English letters. The engineering aspect concerns whether such a design can

efficiently handle the assigned jobs. For instance, the signal-vehicle "a single dot" is assigned to the letter E in order to efficiently encode the most frequently used letter in English into the simplest Morse code. To that extent, the system is designed to minimize human errors that a sender-telegrapher might generate in sending messages like "Elephants eat cheese."

NSAs should be designed to handle any message that behaviorally nonresponsive MCP patients can produce. Consider Cruse et al.'s (2011) toe-imagery task. The task conditions should have been designed in favor of MCP patients, such that their responses would be efficiently manifested by the purported brain activities. Such task conditions include, but are not limited to, the verbal instruction "Imagine wiggling all of your toes and relaxing them." The engineering aspect relates to false negatives, which occur when evidence of an effect is not measured even though the target phenomenon is, in fact, present in the test conditions (Peterson, Cruse et al., 2015, p.591). False negatives, of course, are natural consequences of empirical experiments due to the impossibility of eliminating contingent factors like human error. Nevertheless, I am concerned with the kind of false negatives that are not generated by contingent factors that can be eliminated by conducting tasks repeatedly and more precisely. Let's say that a task is structurally "ill-formed" if it generates this kind of false negative. By analyzing the engineering aspect of the toe-imagery task, I will show that the task structurally permits the possibility of false negatives.

In order to analyze the engineering aspect of the toe-imagery task, we must first distinguish between types of mental action. An essential component of mental-motor action is that a subject mentally executes an instructed mental-motor action from the first-person perspective (i.e., kinesthetic mental-motor action; Ehrsson, Geyer and Naito, 2003). Alternatively, subjects can imagine seeing them-

selves or another person performing an action from an external view (i.e., visual motor imagery), which may be primarily visual in character rather than involving kinesthetic characteristics (Sekiyama, 1983). According to Annett (1995), without specific instructions to perform a kinesthetic mental-motor action, such that when the instruction "Imagine wiggling toes" is given, subjects may either perform a kinesthetic mental-motor action or conceive visual motor imagery. Kinesthetic mental-motor action correlates significantly more strongly with activities in the relevant motor areas than does visual motor imagery, implying that visual motor imagery might not activate the relevant motor areas (Neuper et al., 2005). Hence, Cruse et al.'s toe-imagery task should have been designed in such a way that behaviorally nonresponsive MCP patients are clearly instructed to perform a kinesthetic mental-motor action and that they can clearly comprehend the instruction.

Notice that 3 healthy controls (out of 12) in Cruse et al.'s (2011) toe-imagery task did not show the expected brain activities despite the verbal instructions ("Imagine wiggling all your toes on both feet and relaxing them without making any actual motor movement" and "Concentrate on the way your muscles would feel if you were really performing this movement" (Cruse et al., 2011, p.2098)). I suspect that the false negatives (i.e., 3 healthy controls) were generated because the controls failed to comprehend the instructions properly and conceived the visual motor imagery of wiggling toes instead. Regardless of my speculation's accuracy, a more serious problem follows from the task's verbal instructions.

The toe-imagery task is structurally "ill-formed" because it depends on verbal linguistic instructions. In Cruse et al.'s experiment, it might be that some MCP patients did not show the purported brain activities in the toe-imagery task, not because they were not mini-

mally conscious, but because they could not clearly comprehend the instructions and conceived the visual motor imagery instead. According to Kwiatkowska et al. (2019), 34% of 50 minimally conscious post-comatose patients could not build and had difficulties in reading syntactically complex sentences. Most importantly, the instructions of the toe-imagery task consist of sentences that are syntactically far more complex than those in Kwiatkowska et al.'s experiment. Moreover, in general, MCP patients' brain responses to heard words are weaker in terms of power than those of healthy controls (Nigri et al., 2017). In a nutshell, the capacity to comprehend syntactically complex spoken sentences can be dissociated from minimal consciousness. Thus, the toe-imagery task structurally permits the possibility of false negatives.

The same problem generalizes to active paradigms. Tasks that involve mental-motor actions (e.g., the toe-imagery task, the tennis-imagery task, the home-walking task, etc.) essentially depend on subjects' capacity to perform kinesthetic mental-motor actions in response to verbal instructions. Naci and Owen (2013) and Lule et al.'s (2013) target-counting tasks also depend on verbal instructions similar to those in the toe-imagery task.[6] Given that these tasks exhaust those I have ever come across in the literature, I claim that active paradigm tasks in general are structurally "ill-formed."

Tasks in the passive paradigm are in general structurally "ill-formed" as well, because they require a relatively high degree of attention to the given (particularly auditory) stimulations. To explain the problem, we must first understand that traditional behavioral assessments are designed in accordance with the subcategorization of

[6] Among passive paradigm tasks, Naci et al.'s (2014) and Laforge et al.'s (2020) movie-watching tasks entail a similar problem because they depend on the subjects understanding the linguistic narratives of the movies.

the minimally conscious state (MCS) into MCS- (i.e., patients only show nonreflex behavior such as visual pursuit, localization of noxious stimulation, and/or contingent behavior) and MCS+ (i.e., patients show command following Bruno et al., 2012, p.1087). MCS- can be further distinguished in terms of various degrees of capacity to pay attention. For example, the capacity to track an object with the eyes is a mark of degrees of attention in that eye tracking requires an executive function, namely a combination of voluntary saccadic and smooth pursuit eye movement (Ansell, 1995). It is worth noting that a post-comatose patient's having a low degree of attention (e.g., being able to perform eye tracking relatively unstably and for a short time) can still serve as a meaningful sign of minimal consciousness and a prognostic figure for rehabilitation (Ansell, 1993). In short, traditional behavioral assessments can capture the very minimal degree of MCS-.

The odd-ball task is structurally "ill-formed" because the P300b response can be dissociated from the very minimal degree of MCS-. In King et al.'s (2013) experiment, only 52% of 23 conscious controls with brain injuries showed the P300b response. The number of false negatives (i.e., the remaining 48% of controls) is not negligible because it seems that the number cannot be reduced simply by conducting the task repeatedly and precisely. That is, it seems that these false negatives were generated because the conscious controls with brain injuries could not pay strong attention to global oddballs. Moreover, there are relevant experiments suggesting that stimulations like global oddballs require a relatively high degree of attention. In experiments with autistic and schizophrenic patients, such patients demonstrated reduced responses to stimulations similar to global oddballs (Novick et al., 1980; Kärgel et al., 2016). If the capacity to respond to stimulations like global oddballs can be dissociated from

the relevant executive functions like working memory (thus, the very minimal degree of MCS-), then the odd-ball task structurally permits the possibility of false negatives.

This problem with global oddballs applies to passive paradigm tasks in general. They all depend on similar stimulations; Naci et al.'s (2014) and Laforge et al.'s (2020) movie-watching tasks require that subjects are paying attention to (and understanding) the narratives of the movies. Consequently, passive paradigms in general are structurally "ill-formed."

It is worth noting that, structurally, no task can completely avoid the possibility of false negatives, because the absence of information about minimal consciousness is not evidence of the absence of minimal consciousness. Nonetheless, my analysis of the engineering aspect of the relevant tasks shows that such tasks are specifically designed to test MCP patients with the "right" kind of capacities. In other words, my analysis indicates that we need a new NSA task that does not depend on language comprehension or the capacity to recognize violations of temporal regularities.

4. A Proposal: The Misrepresentation Task

I argued that NSAs in their current form are structurally "ill-formed" because they depend on capacities that can be dissociated from minimal consciousness, such as the capacity to comprehend syntactically complex sentences and the capacity to pay attention to stimulations like global oddballs. Recall that a fundamental problem of traditional behavioral assessments is that they depend on the capacity to perform overt behaviors, which can be dissociated from minimal consciousness. It turns out that NSAs raise a similar fundamental problem.

Below, I propose a task for an NSA that is structurally "well-formed" in the sense that it depends on neither language comprehension nor a degree of attention to global oddballs.

An NSA task is diagnostically reliable if it is informationally and structurally reliable. As I explained, the relevant tasks satisfy the former, but are problematic with respect to the latter. I therefore propose an informationally reliable and structurally "well-formed" task, namely, a misrepresentation task, which consists of two parts: a control task and a melting-wax task:

Control task

> Show a subject lukewarm water droplets being sprayed on an instructor's hands, and then spray lukewarm water droplets on the subject's hands (in such a way that the subject sees it).

Melting-wax task

> Show the subject fake melting wax being dropped on the instructor's hands (with the instructor making a facial expression of pain), and then spray lukewarm water on the subject's hands (in such a way that the subject sees the water drops as melting wax drops).

Noxious hot (46°C) stimulation produces localized activities in prefrontal areas (Tracey et al., 2000). I claim that activities in these areas, if observed in the melting-wax task but not in the control task, can be taken as diagnostically reliable neural signals. It is easy to see why the misrepresentation task is structurally "well-formed." It relies neither on language comprehension nor on the capacity to pay attention to an auditory sequence. The task tests whether a subject can misrepresent a non-noxious tactile-stimulus as a noxious tactile-stimulus. As far as a subject has the capacity to misconceptualize lukewarm water as melting wax, the subject can pass the misrepresentation task.

In order to analyze the informational aspect of the misrepresentation task, a distinction between two types of placebo/nocebo responses needs to be discussed. Benedetti et al. (2003) distinguish between placebo/nocebo responses by conditioning and expectation, where the former concerns unconscious processes such as hormone secretion and the latter concerns conscious processes such as pain or motor performance. The misrepresentation task does not involve conditioning: it does not expose a subject to repeated stimulus-behavioral patterns, nor does it reward/punish the subject for behaving in a particular way in response to certain stimulations. Rather, the misrepresentation task tests whether a subject can form an expectation of the forthcoming "noxious" stimulus. According to Colloca and Benedetti (2009), observational social learning produces placebo/nocebo responses by expectation. The purported brain activity in the misrepresentation task is a mark of a nocebo response due to expectation produced by social learning (i.e., the instructor displays a painful facial expression when the tactile stimulus is paired with melting wax). Most importantly, placebo/nocebo responses by expectation generally require executive functions (Benedetti, Carlino and Pollo, 2011, p.239). Thus, if localized activities in prefrontal areas are observed in the melting-wax task but not in the control task, then the best explanation for the activities is the one that appeals to minimal consciousness. In a nutshell, the misrepresentation task is informationally reliable because P(nocebo effect by expectation|activities in prefrontal areas) is greater than P(nocebo effect by conditioning|activities in prefrontal areas).

Notice that although the misrepresentation task is structurally "well-formed" in the sense that it does not depend on either language comprehension or sufficient attention to notice global oddballs, it might still structurally generate false positives. Placebo responses of patients with dementia of the Alzheimer's type are reduced or totally

lacking (Benedetti, Carlino and Pollo, 2011, p.349). So, it is possible that an MCP patient would have the same problem in forming a nocebo response by expectation. I am not proposing the misrepresentation task as a non-false-positive-generating task, but as a supplement to NSAs in their current form, which relies on a different cognitive capacity. Consider traditional behavioral assessments, where the relevant tasks appeal to various types of cognitive capacities, including eye tracking, automated pupillometry (Vassilieva et al., 2019), and functional object-use (Sun et al., 2018). Likewise, the misrepresentation task is one way of diversifying the types of cognitive capacities to which NSAs appeal in testing minimal consciousness in behaviorally nonresponsive MCP patients.

5. Conclusion

The current paper demonstrated that NSAs in their current form are informationally reliable but structurally "ill-formed." They are informationally reliable because the relevant tasks depend on strong correlations between cognitive capacities and the corresponding brain areas. However, they are structurally "ill-formed" because they essentially depend on language comprehension or stimulations like global oddballs as diagnostic means. The primary aim of NSAs is to test for minimal consciousness in behaviorally nonresponsive post-comatose patients because the minimally conscious state can be dissociated from the capacity to perform overt behavior. Given that language comprehension and global oddball recognition can be dissociated from minimal consciousness, I proposed a task that does not involve such capacities and is informationally reliable, namely the misrepresen-

tation task. Consequently, this paper not only reveals the structural limitations of NSAs, but also attempts to diversify the diagnostic means of NSAs.

Acknowledgments. I thank an anonymous reviewer for this journal for comments that prompted important clarifications in the penultimate draft. I am also deeply grateful to Carrie Figdor for constructive feedback on a previous draft.
This work was supported by a research promotion program of SCNU.

Bibliography

Andrews, K., Murphy, L., Munday, R. and Littlewood, C., 1996. Misdiagnosis of the vegetative state: retrospective study in a rehabilitation unit. *The British Medical Journal* [Online], 313(7048), pp.13–16. https://doi.org/10.1136/bmj.313.7048.13.

Annett, J., 1995. Motor imagery: Perception or action? *Neuropsychologia* [Online], 33(11), pp.1395–1417. https://doi.org/10.1016/0028-3932(95)00072-B.

Ansell, B.J., 1993. Slow-to-recover patients: Improvement to rehabilitation readiness: *Journal of Head Trauma Rehabilitation* [Online], 8(3), pp.88–98. https://doi.org/10.1097/00001199-199309000-00011.

Ansell, B.J., 1995. Visual tracking behavior in low functioning head-injured adults. *Archives of Physical Medicine and Rehabilitation* [Online], 76(8), pp.726–731. https://doi.org/10.1016/S0003-9993(95)80526-5.

Bekinschtein, T.A. et al., 2009. Neural signature of the conscious processing of auditory regularities. *Proceedings of the National Academy of Sciences* [Online], 106(5), pp.1672–1677. https://doi.org/10.1073/pnas.0809667106.

Benedetti, F., Carlino, E. and Pollo, A., 2011. How Placebos Change the Patient's Brain. *Neuropsychopharmacology* [Online], 36(1), pp.339–354. https://doi.org/10.1038/npp.2010.81.

Benedetti, F., Pollo, A. et al., 2003. Conscious Expectation and Unconscious Conditioning in Analgesic, Motor, and Hormonal Placebo/Nocebo Responses. *The Journal of Neuroscience* [Online], 23(10), pp.4315–4323. https://doi.org/10.1523/JNEUROSCI.23-10-04315.2003.

Boly, M. et al., 2011. Preserved Feedforward But Impaired Top-Down Processes in the Vegetative State. *Science* [Online], 332(6031), pp.858–862. https://doi.org/10.1126/science.1202043.

Bruno, M.-A. et al., 2012. Functional neuroanatomy underlying the clinical subcategorization of minimally conscious state patients. *Journal of Neurology* [Online], 259(6), pp.1087–1098. https://doi.org/10.1007/s00415-011-6303-7.

Cohen, M.A., Ortego, K., Kyroudis, A. and Pitts, M., 2020. Distinguishing the Neural Correlates of Perceptual Awareness and Postperceptual Processing. *The Journal of Neuroscience* [Online], 40(25), pp.4925–4935. https://doi.org/10.1523/JNEUROSCI.0120-20.2020.

Colloca, L. and Benedetti, F., 2009. Placebo analgesia induced by social observational learning. *Pain* [Online], 144(1), pp.28–34. https://doi.org/10.1016/j.pain.2009.01.033.

Cruse, D. et al., 2011. Bedside detection of awareness in the vegetative state: a cohort study. *The Lancet* [Online], 378(9809), pp.2088–2094. https://doi.org/10.1016/S0140-6736(11)61224-5.

Di, H. et al., 2008. Neuroimaging activation studies in the vegetative state: predictors of recovery? *Clinical Medicine* [Online], 8(5), pp.502–507. https://doi.org/10.7861/clinmedicine.8-5-502.

Ehrsson, H.H., Geyer, S. and Naito, E., 2003. Imagery of Voluntary Movement of Fingers, Toes, and Tongue Activates Corresponding Body-Part-Specific Motor Representations. *Journal of Neurophysiology* [Online], 90(5), pp.3304–3316. https://doi.org/10.1152/jn.01113.2002.

van Erp, W.S., Aben, A.M.L. et al., 2019. Unexpected emergence from the vegetative state: delayed discovery rather than late recovery of consciousness. *Journal of Neurology* [Online], 266(12), pp.3144–3149. https://doi.org/10.1007/s00415-019-09542-3.

van Erp, W.S., Lavrijsen, J.C. et al., 2015. The Vegetative State: Prevalence, Misdiagnosis, and Treatment Limitations. *Journal of the American Medical Directors Association* [Online], 16(1), 85.e9–85.e14. https://doi.org/10.1016/j.jamda.2014.10.014.

Garrido, M.I., Friston, K.J. et al., 2008. The functional anatomy of the MMN: A DCM study of the roving paradigm. *NeuroImage* [Online], 42(2), pp.936–944. https://doi.org/10.1016/j.neuroimage.2008.05.018.

Garrido, M.I., Kilner, J.M. et al., 2009. Repetition suppression and plasticity in the human brain. *NeuroImage* [Online], 48(1), pp.269–279. https://doi.org/10.1016/j.neuroimage.2009.06.034.

Kärgel, C. et al., 2016. The effect of auditory and visual training on the mismatch negativity in schizophrenia. *International Journal of Psychophysiology* [Online], 102, pp.47–54. https://doi.org/10.1016/j.ijpsycho.2016.03.003.

King, J. et al., 2013. Single-trial decoding of auditory novelty responses facilitates the detection of residual consciousness. *NeuroImage* [Online], 83, pp.726–738. https://doi.org/10.1016/j.neuroimage.2013.07.013.

Kondziella, D. et al., 2016. Preserved consciousness in vegetative and minimal conscious states: systematic review and meta-analysis. *Journal of Neurology, Neurosurgery & Psychiatry* [Online], 87(5), pp.485–492. https://doi.org/10.1136/jnnp-2015-310958.

Kwiatkowska, A., Lech, M., Odya, P. and Czyżewski, A., 2019. Post-comatose patients with minimal consciousness tend to preserve reading comprehension skills but neglect syntax and spelling. *Nature Scientific Reports* [Online], 9(1), p.19929. https://doi.org/10.1038/s41598-019-56443-6.

Kycia, R., 2021. Information and brain. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* [Online], (70), pp.45–72. Available at: <https://zfn.edu.pl/index.php/zfn/article/view/514> [visited on 6 December 2022].

Laforge, G., Gonzalez-Lara, L.E., Owen, A.M. and Stojanoski, B., 2020. Individualized assessment of residual cognition in patients with disorders of consciousness. *NeuroImage: Clinical* [Online], 28, p.102472. https://doi.org/10.1016/j.nicl.2020.102472.

Li, T. et al., 2022. Brain information processing capacity modeling. *Scientific Reports* [Online], 12(1), p.2174. https://doi.org/10.1038/s41598-022-05870-z.

Lulé, D. et al., 2013. Probing command following in patients with disorders of consciousness using a brain–computer interface. *Clinical Neurophysiology* [Online], 124(1), pp.101–106. https://doi.org/10.1016/j.clinph.2012.04.030.

Monti, M.M. et al., 2010. Willful Modulation of Brain Activity in Disorders of Consciousness. *New England Journal of Medicine* [Online], 362(7), pp.579–589. https://doi.org/10.1056/NEJMoa0905370.

Naccache, L., 2018. Minimally conscious state or cortically mediated state? *Brain* [Online], 141(4), pp.949–960. https://doi.org/10.1093/brain/awx324.

Naci, L., Cusack, R., Anello, M. and Owen, A.M., 2014. A common neural code for similar conscious experiences in different individuals. *Proceedings of the National Academy of Sciences* [Online], 111(39), pp.14277–14282. https://doi.org/10.1073/pnas.1407007111.

Naci, L. and Owen, A.M., 2013. Making Every Word Count for Nonresponsive Patients. *JAMA Neurology* [Online], 70(10), pp.1235–1241. https://doi.org/10.1001/jamaneurol.2013.3686.

Neuper, C., Scherer, R., Reiner, M. and Pfurtscheller, G., 2005. Imagery of motor actions: Differential effects of kinesthetic and visual–motor mode of imagery in single-trial EEG. *Cognitive Brain Research* [Online], 25(3), pp.668–677. https://doi.org/10.1016/j.cogbrainres.2005.08.014.

Nigri, A. et al., 2017. The neural correlates of lexical processing in disorders of consciousness. *Brain Imaging and Behavior* [Online], 11(5), pp.1526–1537. https://doi.org/10.1007/s11682-016-9613-7.

Noh, H., 2018. No-report Paradigmatic Ascription of the Minimally Conscious State: Neural Signals as a Communicative Means for Operational Diagnostic Criteria. *Minds and Machines* [Online], 28(1), pp.173–189. https://doi.org/10.1007/s11023-017-9433-6.

Noh, H., 2022. Behavioral vs. Neural Methods in the Treatment of Acutely Comatose Patients. *Ramon Llull Journal of Applied Ethics* [Online], 1(13), pp.245–258. https://doi.org/10.34810/rljaev1n13Id398703.

Novick, B., Vaughan, H.G., Kurtzberg, D. and Simson, R., 1980. An electrophysiologic indication of auditory processing defects in autism. *Psychiatry Research* [Online], 3(1), pp.107–114. https://doi.org/10.1016/0165-1781(80)90052-9.

Owen, A.M. et al., 2006. Detecting Awareness in the Vegetative State. *Science* [Online], 313(5792), p.1402. https://doi.org/10.1126/science.1130197.

Peterson, A. and Bayne, T., 2018. Post-comatose disorders of consciousness. In: R.J. Gennaro, ed. *The Routledge Handbook of Consciousness* [Online], *The Routledge Handbooks in Philosophy*. New York: Taylor & Francis, pp.351–365. https://doi.org/10.4324/9781315676982.

Peterson, A., Cruse, D. et al., 2015. Risk, diagnostic error, and the clinical science of consciousness. *NeuroImage: Clinical* [Online], 7, pp.588–597. https://doi.org/10.1016/j.nicl.2015.02.008.

Pfurtscheller, G. and Neuper, C., 1997. Motor imagery activates primary sensorimotor area in humans. *Neuroscience Letters* [Online], 239(2-3), pp.65–68. https://doi.org/10.1016/S0304-3940(97)00889-6.

Pulvermüller, F., 2005. Brain mechanisms linking language and action. *Nature Reviews Neuroscience* [Online], 6(7), pp.576–582. https://doi.org/10.1038/nrn1706.

Rac-Lubashevsky, R. and Kessler, Y., 2019. Revisiting the relationship between the P3b and working memory updating. *Biological Psychology* [Online], 148, p.107769. https://doi.org/10.1016/j.biopsycho.2019.107769.

Raposo, A., Moss, H.E., Stamatakis, E.A. and Tyler, L.K., 2009. Modulation of motor and premotor cortices by actions, action words and action sentences. *Neuropsychologia* [Online], 47(2), pp.388–396. https://doi.org/10.1016/j.neuropsychologia.2008.09.017.

Rohaut, B., Eliseyev, A. and Claassen, J., 2019. Uncovering Consciousness in Unresponsive ICU Patients: Technical, Medical and Ethical Considerations. *Critical Care* [Online], 23(1), p.78. https://doi.org/10.1186/s13054-019-2370-4.

Schnakers, C. et al., 2009. Diagnostic accuracy of the vegetative and minimally conscious state: Clinical consensus versus standardized neurobehavioral assessment. *BMC Neurology* [Online], 9(1), p.35. https://doi.org/10.1186/1471-2377-9-35.

Sekiyama, K., 1983. Mental and physical movements of hands: Kinesthetic information preserved in representational systems. *Japanese Psychological Research* [Online], 25(2), pp.95–102. https://doi.org/10.4992/psycholres1954.25.95.

Shannon, C.E., 1948. A mathematical theory of communication. *The Bell System Technical Journal* [Online], 27(July 1928), pp.379–423. https://doi.org/10.1145/584091.584093.

Sun, Y. et al., 2018. Personalized objects can optimize the diagnosis of EMCS in the assessment of functional object use in the CRS-R: a double blind, randomized clinical trial. *BMC Neurology* [Online], 18(1), p.38. https://doi.org/10.1186/s12883-018-1040-5.

Teasdale, G. and Jennett, B., 1974. Assessment of coma and impaired consciousness. *The Lancet* [Online], 304(7872), pp.81–84. https://doi.org/10.1016/S0140-6736(74)91639-0.

Tracey, I. et al., 2000. Noxious hot and cold stimulation produce common patterns of brain activation in humans: a functional magnetic resonance imaging study. *Neuroscience Letters* [Online], 288(2), pp.159–162. https://doi.org/10.1016/S0304-3940(00)01224-6.

Vassilieva, A. et al., 2019. Automated pupillometry to detect command following in neurological patients: a proof-of-concept study. *PeerJ* [Online], 7, e6929. https://doi.org/10.7717/peerj.6929.

Wang, F. et al., 2019. Detecting Brain Activity Following a Verbal Command in Patients With Disorders of Consciousness. *Frontiers in Neuroscience* [Online], 13, p.976. https://doi.org/10.3389/fnins.2019.00976.

Weaver, W., 1953. Recent Contributions to the Mathematical Theory of Communication. *ETC: A Review of General Semantics* [Online], 10(4), pp.261–281. Available at: <https://www.jstor.org/stable/42581364> [visited on 20 January 2023].

Is information ontological or physical, or is it perhaps something else? Some remarks on Krzanowski's approach to the concept of information

Łukasz Mścisławski
Wrocław University of Science and Technology, Poland

Abstract

As one may have noticed, the title of this paper is somewhat provocative. We found Roman Krzanowski's (2020a,b,c; 2022) proposed approach to the problem of information very intriguing. Our aim here is to highlight some advantages when it comes to answering some fundamental questions in the philosophy of physics and metaphysics, as well as the philosophy of information and computer science. This issue is of great importance, so we propose that the introduction of some subtle distinctions between ontological and epistemological information can be regarded as being analogous to G.F.R. Ellis's analyses of the passage of time in his concept of the Crystallizing Block Universe (Ellis and Goswami, 2012). This analogy could be useful when further studying the relations between different types of information. We also suggest some subjects for further study, ones where Krzanowski's proposal could serve as a very solid foundation for examining traditional metaphysical issues by combining classical philosophical doctrines with the new approach.

Keywords

physical information, ontology, physics, philosophy of information.

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 147–169

Introduction

This article primarily aims to concisely present Roman Krzanowski's approach to the concept of *physical information.* This concept is perceived here as a kind of "intermediate concept" that can act as a bridge to the development of his concept of *ontological information* (cf. Krzanowski, 2020a,b,c; and especially Krzanowski, 2022),and can be regarded as a special case of the latter. Indeed, we believe that the latter concept inherits many characteristics from the former, but not all the issues have been solved.[1] This situation implies that we are dealing with a truly fundamental issue here, namely the fundamentality of the concept of information. Furthermore, this paper also aims to discuss some of this concept's interesting properties and present possible avenues for further developing this interesting proposal. It is well known that there is enormous interest in the philosophical aspects of the concept of information (e.g., Adriaans and Benthem, 2008; Burgin, 2010; Floridi, 2011; Dodig-Crnkovic and Burgin, 2019), but we feel that Krzanowski's proposal is particularly attractive with some great research potential, especially for studying physical reality. It would therefore be helpful, in our opinion, to position this concept within the vast array of philosophical problems related to the notion of information, as well as introduce a few distinctions to allow us to order the discourse.

The first distinction we would like to introduce, because we consider it important, is the distinction between two perspectives for the notion of information:

(A) Ontological; and

[1] It is also worth mentioning that in some works, Krzanowski seems to use these two terms (i.e., physical and ontological information) interchangeably (cf. Krzanowski, 2020b).

(B) Epistemological.

These two areas of research are by no means mutually exclusive, but they pose slightly different kinds of questions. Krzanowski (2020c, pp.37–38; 2022, pp.123–151) presented an analogous distinction.[2] For the purposes of this paper, however, they can be characterized as follows: The *ontological* (A) perspective mainly poses questions about what something is, how it exists, what its inherent properties are, and so on. It is worth noting here that within this research area, it is possible to pose a question about the general structure of reality (i.e., its ontology and the "location" of information within it) and thus its ontological status. Such a general perspective makes it possible to regard information as something tangible, so there is no need to reify it, although we cannot exclude such a possibility. This perspective is of particular interest to Krzanowski and therefore also to this paper.

The *epistemological* (B) perspective, on the other hand, poses questions that are typical of epistemology, such as how something can be cognitively grasped and whether and how it relates to issues of knowledge, truth, cognition, and so on.

The second distinction that naturally arises when considering the notion of information relates to two approaches:

(I) Qualitative; and
(II) Quantitative.

It is plain to see that these two approaches can be combined with the above perspectives. We could say that the epistemological perspective

[2] A slightly different version is also presented in (Krzanowski, 2020b), where a distinction between abstract and concrete information is introduced. There is a little more on this subject below.

[2]As a rule in this paper, ontology is not understood as a network of concepts and relations between them, as is the case in some formal disciplines.

(B) is compatible with both approaches (I and II). Similarly, it is clear that perspective A must be compatible with approach I. Nevertheless, some interesting issues arise:

1. Can every characteristic of I be expressed in the form of II? It seems that the answer to this question is by no means obvious, and trying to provide one may be better regarded as a starting point for some interesting and deep research into addressing the general problem of relating the two approaches.
2. Is it possible to combine perspective A with approach II for the concept of information? This is not something that could be achieved by simply declaring that "information is quantitative only". It seems that any serious attempt to answer this sort of question would possibly require connecting what we call *information* with mathematical structures. Perhaps with regard to the ontological status of the latter, a perspective along the lines of mathematical Platonism should be included here, at least to some extent.

It is also worth noting another difficulty with approaches of type II, which can be formulated as follows: *What is that* which is being quantitatively represented? This question becomes all the more pertinent when one accounts for Burgin's (2011, p.349) observation, where the *information* is something and the *what* is a measure imposed on it.[3] Of course, it can always be argued that we are using a projective definition here, but this would only serve to cut off the discussion instead of resolving the difficulty. In fact, this would be an ineffective

[3] An analogy could be drawn with the subtle distinction between space and distance as an imposed measurement of it. It is otherwise surprising how many analogies from analyses of foundational physics are applicable to considerations of the concept of information.

ploy, because over 30 quantitative accounts of information can be given (Burgin, 2010, pp.131–133). Hence, two fundamental questions arise here: Firstly, which of these approaches should be adopted as a starting point and why? Secondly, are these two approaches related to each other, and if so, how?

Thus, if we assume that quantitative approaches do not provide a good basis for considering the concept of information, then it seems reasonable to attempt tackling this issue in a different way. Quantitative approaches (II) seem to be strongly linked to the epistemological perspective of research (B), so we should perhaps place more emphasis on the qualitative approach (I) and try to look at it more from an ontological perspective (A). Roman Krzanowski's conception undoubtedly falls within such an area of research. It represents what, at least in a sense, Floridi would call *information for reality* (Floridi, 2011, pp.30–31).[4] The difference between Floridi's account and Krzanowski's proposal basically lies in how the latter approach concerns *physical reality*. This observation requires further elaboration, because the problems concerning the relationships between physics and information are widely discussed. It should be noted, however, that these propositions are more quantitative in nature and approached from an epistemological perspective (B), which in a way seems counter to Krzanowski's conception.

In the following discussion, the concept of physical information proposed by Krzanowski will therefore be presented, and its basic properties will be discussed. Potential avenues for future research will also be discussed.

---

[4] We here use the label "information for reality" rather than Floridi's concept of information. In Floridi's work (cited above) there is an inconsistency between the definition and exactly using the notion of information in the case of "information for reality". We are very grateful to anonymous reviewer for bringing this to our attention.

The concept of physical information

"So Professor Isham, what is a thing?"
"We can't say what is a thing, but you can say what is not."
"What is not?"
"Not what people think it is"
(Medeiros, 2005)

The discussion quoted above may seem humorous, but this is a common situation when fundamental concepts become the subject of research. In our opinion, research into the concept of information belongs to such research. Thus, the next step in facilitating a deeper exploration of the notion of *physical information* is to follow the example of Krzanowski in introducing a distinction between *abstract information* and *concrete information* (cf. Krzanowski, 2020b, p.2).

The concept of *abstract information* (IA) relates to some kind of cognitive activity, and based on Krzanowski's work, its main features can be expressed as follows (cf. Krzanowski, 2020b):

(IA1) It is some cognitive agent's interpretation of physical stimuli, which may be a signal, the state of physical system, or some other physical phenomenon.
(IA2) It exists for a cognitive agent, or it is at least relative to some agent, so it is agent-relative or ontologically subjective.
(IA3) It has meaning for a cognitive agent.
(IA4) The notion of a cognitive agent is understood here in a very broad sense, such that it may be human, another biological system, or some artificially intelligent system.
(IA5) The existence of IA indicates the presence of an abstract notion somewhere outside of space and time.

When discussing the concept of information, the IA concept plays an important role. It seems that almost all the quantitative formulations

we mentioned earlier can be assigned to this category of information, because they are in a sense imposed on physical reality by the cognitive subject. Moreover, due to the research successes of physics, which employs mathematical methods to a large extent, there is considerable temptation to narrow any discussion about the concept of information to references to physical reality.

Nevertheless, Krzanowski's concept of *physical information* refers to the concept of information using the term *concrete information* (IC). He refers to an extensive list of authors whose views converge in this respect (e.g., Turek, 1978; von Weizsäcker, 1982; Nagel, 2012; Dodig-Crnkovic, 2013; Heller, 2014; Rovelli, 2016; Wilczek, 2016; Davies, 2020), to name but a few). He also seems to be guided by an opposition to attributing only the features of IA to information in general (cf. Krzanowski, 2020a,c). This triggers a need to introduce a different approach to the concept of information, a more qualitative one (IC). Thus, the fundamental features of IC can be described as follows (cf. Krzanowski, 2020b, p.2):

(IC0) IC exists in space and time (i.e., spacetime) as a physical object, which is why it is called *concrete*.
(IC1) With reference to IC0, IC is a physical phenomenon, so it exists objectively and is not relative to anything.
(IC2) IC has no intrinsic meaning.
(IC3) IC is, in a sense, responsible for the organization of the physical world.
(IC4) IC's existence implies existence in the physical world, somewhere in the space-time continuum.

The main goal behind introducing the IC concept is, according to Krzanowski, a hope that it may unify multiple quantitative approaches, at least at a conceptual level, or establish some order among the

multiplicity of formulations. It therefore seems reasonable to conclude that his concept of *physical information* refers to *concrete information* that is "*associated with*" the physical level of the organisation of matter. This statement requires some elaboration and clarification, however. It is also worth emphasizing that only with the concept of concrete information is there at least some way to use it within the general discourse about information, including possibly regarding information as meta-physical.

One can easily see how some may object to various properties of this concept, but since we engage in a broader discussion of these properties later in this paper, it is more appropriate for now to continue presenting further key features of Krzanowski's proposal.

Undoubtedly, one of the most important features of *concrete information* is its *objective existence* (IC1). As Krzanowski puts it, this means it exists as a physical phenomenon or object, independently of any observing agent (cf. Krzanowski, 2020b, pp.4–5).

The second feature emphasized by Krzanowski is IC's lack of intrinsic meaning (IC2), referring to how meaning is derived from an observed reality (e.g. a physical object, phenomenon, etc.) by a cognitive agent. Since we are here discussing physical reality in itself, this means it has no meaning of its own. This reality can be interpreted from many points of view, but the procedure of deriving meaning is actually a shift into the realm of *abstract information* (cf. Krzanowski, 2020b, pp.5–6).

The presentation of the third feature is very, possibly even hopelessly, difficult. As Krzanowski emphasizes, any discussion of the concept of information becomes interwoven with notions such as *form*, *structure*, *object*, and so on. Another fundamental problem here emerges when one tries to understand what it means for *information* to be *associated* with concepts like form and structure. To some extent,

however, we can say that IC is in some way responsible for the organization of matter. This statement needs clarification, which Krzanowski provides by addressing the question of whether IC can be considered a physical phenomenon (cf. Krzanowski, 2020b, pp.3–4). More specifically, he describes *physical information* binding with physical reality and mathematics (or mathematical structures) as follows:

(PhI1) Physical information, as an inheritor of concrete information, is described as a physical phenomenon. It should be highlighted that Krzanowski gives a very special meaning to this statement, because physical information being a physical phenomenon implies that this special type of information is an irreducible aspect of physical reality. In a way, it recognizes something—whether it be the form, structure, or organization of some entity—as a purely physical phenomenon in itself.

(PhI2) Physical information exhibits properties that can be attributed to physical entities, namely that it:

- (PhI2.a) is observable;
- (PhI2.b) is ontologically objective;
- (PhI2.c) can be manipulated;
- (PhI2.d) has no intrinsic meaning; and
- (PhI2.e) can be quantified or measured.

(PhI3) Physical information is not a mathematical or physical structure, thus preventing it from being considered as part of the realm of mathematical or physical structures, something that could easily lead to referring to such structures rather than to the (physical) information itself. However, where physical reality exists, there must also exist physical information (cf. Krzanowski, 2020b, pp.3–4).

What is also significant is how Krzanowski does not insist that his concept is well-defined without any ambiguities. In contrast, he emphasizes that in any serious analysis of the concept of information and its relation to the physical world, there will be ambiguities. Moreover, such ambiguity is characteristic of how physical reality and information are related. Like Krzanowski, we believe that remaining at a more or less descriptive level is unavoidable when addressing such a subtle and intangible issue, thus excluding any narrow perspective that someone could subjectively call "sensible" (cf. Krzanowski, 2020b, p.6).

Remarks and potential avenues for further development

Starting with some additional remarks, we refer to the epistemological perspective for information and quantitative approaches. We begin with the obvious statement that any cognitive act is possible if and only if a cognitive agent can cognize something. This leads to the following statement: In the reality in which we are able to cognize, there exist entities such as cognitive agents and entities that they can cognize.[5] If their existence is long and stable enough, then an act of cognizance is possible. We believe that this situation strongly suggests that some structures must exist in physical reality, that there are cognitive agents at a physical level, and that these agents can perceive the structures in some way. This then leads us to the conclusion that the structures of physical reality precede acts of cognition. Thus, it

[5] We do not want to start a battle here about whether they are able to build knowledge about their reality but rather state that they cognize without discussing what knowledge is. We simply need to assume that there are various stimuli that agents can perceive and react to in a certain way.

seems that all quantitative approaches and epistemological perspectives for research into the concept of information come secondary to one very fundamental fact: There are physical structures. We understand physical structure in a very broad manner as something that can be in some way distinguished from its background. For example, in this sense, even an elementary particle can be regarded as a structure, because it can be viewed as an excitation (or an excited state) of the quantum field. We are not suggesting here that it should be understood as something that is separated from its background but rather that we are allowed to say that *there is a physical structure if there is any differentiation in a considered physical reality*, regardless of whether we can describe this differentiation mathematically or not.[6] In our opinion, to answer questions about how this is possible, one of the most obvious ways would be to point out how the laws of physics tell matter how to behave. However, we here encounter extremely difficult questions about the relations between matter, information, and the laws of physics. Nevertheless, we would like to emphasize that much depends on how we understand the laws of physics. The first option would be to define the laws of physics as part of our description of the regularities in physical reality (PLE).[7] This means that human beings observing these regularities of physical reality act as cognitive agents trying to express these regularities using mathematical structures. However, this returns us to the epistemological perspective. The second option lies in the definition of the laws of physics (PLO), such that we could assume that a kind of Platonic realm for mathematical structures exists, and a portion of those structures govern and shape

[6] There are two interesting properties of such an approach: 1) We are completely free in referring to physical objects as parts of wider structures, and 2) we are free to regard physical objects as entities of internal structure, even though it may be infinitely complex.

[7] We treat physical laws here as scientific laws.

matter (e.g., Heller, 1998; Penrose, 2006; see also Grygiel, 2022). Such a possibility immediately opens a door to the ontological perspective,[8] and within this context, we would like to emphasize the importance of the ontological perspective in research into the concept of information (cf. Krzanowski, 2020c, p.53). Nevertheless, we are interested in attempting to answer the question of why regularities in physical reality can exist at all? This question involves the relationships between matter, information, and the laws of physics, and it is all the more difficult because all those concepts are highly problematic. This is exactly why we regard Roman Krzanowski's concept of physical information as such an interesting proposal. It seems to make it possible to at least partly answer Hawking's famous question: *What is it that breathes fire into the equations and makes a universe for them to describe*? (Hawking, 1988, p.174).

R1: Our first remark refers to the feature PhI3 and the possible relations that *physical information* has with physical and mathematical structures. Krzanowski claims that where physical reality exists, there is physical information, and this suggests two possibilities:

(a) Physical information is something inherent in matter, but this solution excludes any further discussion of the laws of physics, the possibility of cognizing physical reality, and so on.[9]

(b) Physical information is s*omehow different from physical reality*, which is the research domain of physicists. Nevertheless, it is

[8] We have to admit here that opening such a door also opens up a Pandora's box of questions about mutual relations, such as the "Platonic" world of mathematical structures, matter, the mind, and so on, but we will skip over this endless discussion here.

[9] We dare to posit that such a solution is unsatisfying and of little interest. However, it still leaves us with unanswered questions: What is matter? Why is it formed in the way it is?

> *somehow associated with it*, albeit with a different ontological status. It reveals itself through the existence of physical structures and the opportunity to cognize physical reality, even with measurement. Thus, we regard this as a strong case for regarding it as a *meta-physical* reality, one that is tangible because it is "responsible for" creating physical structures. This last claim in some way justifies thinking of physical information as something *inherent* to any physical object or phenomenon as an "internal" (meta-physical) constituent of it. Additionally, belonging to the ontological level and existing prior to any cognitive agent, physical information turns out to be more fundamental than any quantitative definition of information, albeit with the caveat contained in R2.

R2: It seems to us that PhI3 suggests that, in some way, the concept of physical information is not to be regarded as a concrete mathematical structure. We would like to point out that this feature of the discussed concept needs further development. We claim that at the moment, the concept of physical information and its relations strongly depend upon what ontological assumptions are made, such as what is assumed to exist, whether there is any kind of metaphysical pluralism, and how particular types of entities (or structures) interrelate. For example, if a kind of Platonic ontological structure of reality is assumed,[10] it could also be assumed that a part of the objectively existing Platonic mathematical world completely models (or causally acts upon and determines) physical reality, so physical structures are merely a material representation of particular mathematical structures. In such a case, physical information would surely be associated with

[10] A good example of such a structure of reality was described by Penrose (2006, pp.18–19).

certain mathematical structures, and this could be considered within any quantitative approach. There are of course easy way to escape this difficulty. More specifically, it suffices to assume that the mathematical structures of a Platonic world do not describe the entirety of physical reality.[11]

R3: Because physical information in some way inherits the features of concrete information, there is some difficulty with IC0, IC4, and PhI1. All these features suggest that physical information is "located" on the Newtonian-like stage of space and time. However, it seems that as Krzanowski describes it, space and time (or spacetime) are independent of physical information. General Relativity, however, describes spacetime (or space and time) as part of physical reality. Hence, this strongly suggests that we should regard physical information as something that is also in some way associated with the structure of spacetime. In other words, it "contains information" for spacetime. This aspect of the concept proposed by Krzanowski gives further backing for regarding physical information as something metaphysical while still being strictly connected to, or associated with, the physical level of the organization of matter.

R4: In all the key works (i.e., Krzanowski, 2020b,c; 2022), there is a lack of any linkage to quantum theories (e.g. quantum mechanics, quantum field theory, etc.). There is also no remark about no-go theorems (e.g. Kochen-Specker theorem), contextuality, and so on. These omissions are a little bit puzzling, but they could be explained in two ways:

[11] Such an example was also described by Penrose (2006, pp.19–21).

(1) The concept of physical information refers to the very physical reality (cf. IC3 and PhI1), and as such, it also refers to the strange quantum realm. There is therefore no need to confer a special status to this realm or any issues connected with the strange quantum features of it.

(2) As with all physical theories, quantum theories are not ultimate theories but rather something through which we try to describe and explain physical reality. In this respect, it is a purely epistemological perspective, while the subject of interest (the concept of physical information) adopts an ontological perspective. However, in this case, a question arises as to whether any suggestions from philosophical research into quantum theories should be considered when exploring the concept of information, particularly for physical information. This matter requires extreme caution, however, because scientific theories tend to evolve relatively quickly, so drawing any far-reaching ontological conclusions from them is a difficult undertaking. We therefore regard this issue as an open question.

R5: By virtue of IC3 and PhI1, the concept of physical information refers to physical reality, but it is not clear whether it is associated with the entirety of physical reality or just particular structures. In the latter case, a question naturally arises about the relations in which particular "physical information" is associated with particular structures. On the other hand, it is precisely this reference to physical reality that allows the concept under discussion to be open to being "contained" in concepts of information, meaning higher levels of organisation of matter, such as chemical, biological, and so on. This opens up a very interesting research area that relates to the possible types of information, their mutual relationships, and particularly the complexity (highly complex,

non-linear, and chaotic systems) of it all. Indeed, Krzanowski (2020b, p.13) recognizes these areas. Another interesting question refers to potential relations with issues connected with computer science and natural computation. It again seems by virtue of the fact that physical information refers to physical reality, it could be included in such analyses. This problem is partially addressed by Krzanowski (2022), but we believe that it warrants further research, especially within the context of computations and relations between physical structures, some of which are very special, such as computing devices[12] and mathematical structures.

R6: If physical information is to be regarded as something that refers to structures in nature (Krzanowski, 2022, pp.86), we should also account for the following issues:

(a) Physical structures are dynamic, so we should try to answer the following question: Are changes in physical structures really also changing the physical information, so it should be regarded as dynamic. Or does physical information contain the "dynamics" of these physical structures? Is it perhaps rather the case that changes in physical reality, caused naturally or otherwise, take place in a manner that is determined by physical information?

(b) Perhaps it is a good idea to regard physical information as something "standing behind" physical structures and their dynamics (changes). One possibility would be to view physical information as a kind of potentiality for creating (physical)

[12] By computing devices, we refer to artificial computing devices like personal computers, while we assume that any natural process can be regarded as a form of computation, so any "natural computing" device is a natural process.

structures.[13] This could be based on the fairly obvious observation that structures exist in nature, and nature tends to create structures, yet the idea needs further research, because the potential to create structures does not necessarily stem from the fact that there has to be structures.[14]

(c) It seems that physical information, by virtue of being "responsible" for manifesting physical structures, makes it possible for epistemic concepts of information to exist.

R7: A subtle issue arises when we question the ontological status of physical information, as well as its genesis. Indeed, there are many opportunities for further research in this area (cf. Krzanowski, 2020b, p.13). There is also a very important question about the causal relations between physical information and the physical reality with which it is connected. Within this context, the problematic relations between matter, physical information, and the laws of physics arise once more (e.g., Davies, 2007).

As has been presented thus far, physical information is described as something that "stands behind" physical structures, while many points suggest that it has no physical character of its own (see also Burgin, 2017). It therefore seems quite natural to treat physical information as being metaphysical or, in other words, ontological information (cf. Krzanowski, 2020b; 2022). Such a move makes it possible

[13] From private correspondence with R. Krzanowski.

[14] It should be noted that this issue was addressed by Czesław Białobrzeski, among others. To explain how it is possible for structures to arise in nature, Białobrzeski adapted the ontological ideas of Nicolai Hartmann and introduced the category of organisation (Polish: *kategoria ustrojowości*), which is responsible for allowing higher layers of reality to arise, as well as a real factor that he called potentiality that is responsible for the state and organisation of a system (cf. Białobrzeski, 1984, pp.243–247; Mścisławski, 2017). This area of research is closely connected to the issue of the relation between physical information and complexity (see R5 above).

to view this information as referring to physical systems rather than being a physical phenomenon in itself. As Krzanowski puts it, ontological information is not something from the Platonic world but rather something that is closely connected with physical reality, with it unveiling itself much like physical phenomena and their properties do (cf. Krzanowski, 2022, p.110). Yet another question arises here, however: How does ontological information relate to the laws of physics? If we assume that we define these laws as PLE, the solution is relatively simple. The real problem arises when we define these laws as PLO, and this represents another potential area for further research. What is also interesting here is that this step also positions information as possibly having two modes of existence, namely concrete and abstract (Krzanowski, 2022, pp.154), so it does not ultimately solve the fundamental difficulties of their relations to spacetime, the problem of causality relations, and the issue of complexity.

We would like to suggest another potential research area that addresses the issue of treating physical information (and ontological information) as being associated with a very special kind of transition, namely the transition from ontological possibility to a concrete physical structure (reality).[15] Is physical information therefore to be regarded as a transition, a kind of "ontological process" that is analogous to the forming of matter in hylomorphism or rather as an analogy of form (cf. Krzanowski, 2020b, p.13)? Or should it be regarded as

[15] This proposal is analogous to Ellis' proposal of understanding "now" as the transition from the future, which is understood as ontological undetermination (uncertainty), to the past, which is understood as epistemological uncertainty (cf. Ellis and Goswami, 2012). In this approach to the concept of physical information, it is also important that we refer to a transition from ontological possibilities to actual emerging physical states of reality, and we are not referring just to information about possibilities (initial possible states) and information about actuality (final actual states). We see a certain similarity between this transition and the proposed mechanism of decoherence for solving the problem of vector state reduction in quantum mechanics (cf. Zurek, 2002).

a kind of description or algorithm for another factor acting on matter?[16] Perhaps further research into the abovementioned transition could shed some light on the links between physical or ontological information and causality. Indeed, these open questions could be a starting point for further research.

In our opinion, all the remarks mentioned above lead us to yet another potential area of research, and the question addressed within it could be formulated as follows: What kind of ontology would be extensive enough to encompass all possible types of beings[17] and existence, such that it could deal with all the complex issues? The situation becomes even more complicated if we also include the issue of virtual beings (e.g., Skowron, 2020) and relations between virtual reality (or realities) and physical reality. It seems that a kind of combined ontology may be needed, such as one based on the proposal of Perzanowski (2016).

Conclusion

We find Krzanowski's proposed concept of physical information very interesting, particularly at a certain stage in his study of the concept of information. While there are many points in which this concept seems to be ambiguous, there are also some interesting areas for possible further research. Thus, we have endeavoured to present the concept and point out the problematic aspects. Most of these could be regarded as potential starting points for further study, despite the

[16] If this is the case, we would rather regard physical information both as a description (data) and a kind of algorithm. It would determine the features of structures, their behaviour (dynamics), and both a physical and ontological manifestation.

[17] As a kind of being (something that exists), we also refer here to structures of any type and kind.

fact that some of these ambiguities were partially clarified in the concept of ontological information presented by Krzanowski (2022). Nevertheless, presenting this concept in light of the issues presented above warrants a separate study.

Acknowledgements. We would like to express our gratitude to the two anonymous reviewers for their careful reading of our manuscript and extremely valuable and helpful comments.

Bibliography

Adriaans, P. and Benthem, J.v., 2008. *Philosophy of Information*. Amsterdam, The Netherlands; Boston: North Holland.

Białobrzeski, C., 1984. *Podstawy poznawcze fizyki świata atomowego*. Wyd. 2 rozszerzone. Warszawa: Państwowe Wydawnictwo Naukowe.

Burgin, M., 2010. *Theory of Information: Fundamentality, Diversity and Unification*. Vol. 1, *World Scientific Series in Information Studies*. Singapore: World Scientific.

Burgin, M., 2011. Information: Concept Clarification and Theoretical Representation. *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society* [Online], 9(2), pp.347–357. https://doi.org/10.31269/triplec.v9i2.284.

Burgin, M., 2017. The General Theory of Information as a Unifying Factor for Information Studies: The Noble Eight-Fold Path. *Proceedings* [Online], 1(3), p.164. https://doi.org/10.3390/IS4SI-2017-04044.

Davies, P.C.W., 2007. The Implications of a Cosmological Information Bound for Complexity, Quantum Information and the Nature of Physical Law. In: C. Claude and G.J. Chaitin, eds. *Randomness and Complexity, From Leibniz to Chaitin* [Online]. World Scientific, pp.69–87. [visited on 26 January 2022].

Davies, P., 2020. *The Demon in the Machine: How Hidden Webs of Information Are Solving the Mystery of Life*. London: Penguin Books.

Dodig-Crnkovic, G., 2013. Alan Turing's Legacy: Info-computational Philosophy of Nature. In: G. Dodig-Crnkovic and R. Giovagnoli, eds. *Computing Nature: Turing Centenary Perspective* [Online], *Studies in Applied Philosophy, Epistemology and Rational Ethics*. Berlin; Heidelberg: Springer, pp.115–123. https://doi.org/10.1007/978-3-642-37225-4_6.

Dodig-Crnkovic, G. and Burgin, M., eds., 2019. *Philosophy and Methodology of Information: The Study of Information in a Transdisciplinary Perspective*, *World scientific series in information studies*, 10. New Jersey: World Scientific.

Ellis, G.F.R. and Goswami, R., 2012. Space Time and the Passage of Time. *arXiv:1208.2611 [gr-qc]* [Online]. Available at: <http://arxiv.org/abs/1208.2611> [visited on 10 June 2014].

Floridi, L., 2011. *The Philosophy of Information*. Oxford: Oxford University Press.

Grygiel, W.P., 2022. A critical analysis of the philosophical motivations and development of the concept of the field of rationality as a representation of the fundamental ontology of the physical reality. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (72), pp.87–108.

Hawking, S., 1988. *A Brief History of Time: From the Big Bang to Black Holes*. New York; Toronto: Bantam Books.

Heller, M., 1998. Czy świat jest matematyczny? *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (23), pp.3–14.

Heller, M., 2014. *Elementy mechaniki kwantowej dla filozofów*. Wyd. 1 w tej edycji. Kraków: Copernicus Center Press.

Krzanowski, R., 2020a. Does Purely Physical Information Have Meaning? A Comment on Carlo Rovelli's Paper: Meaning = Information + Evolution. *arXiv:2004.06716 [physics]* [Online]. Available at: <http://arxiv.org/abs/2004.06716> [visited on 30 November 2021].

Krzanowski, R., 2020b. What Is Physical Information? *Philosophies* [Online], 5(2), pp.9–22. https://doi.org/10.3390/philosophies5020010.

Krzanowski, R., 2020c. Why can information not be defined as being purely epistemic? *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* [Online], (68), pp.37–62. Available at: <https://zfn.edu.pl/index.php/zfn/article/view/494> [visited on 30 November 2021].

Krzanowski, R., 2022. *Ontological Information: Information in the Physical World* [Online]. Vol. 13, *World Scientific series in information studies*. Hackensack, NJ: World Scientific. https://doi.org/10.1142/12601.

Medeiros, J., 2005. Surely you're joking Professor Isham? *I, Science: The Imperial College science magazine*, (3, Winter), p.20.

Mścisławski, Ł., 2017. Między filozofią w nauce a filozofią przyrody: Piękno systemu w konkrecie: zakaz Pauliego a filozofia Czesława Białobrzeskiego. In: P. Polak, W.P. Grygiel and J. Mączka, eds. *Oblicza filozofii w nauce: księga pamiątkowa z okazji 80. urodzin Michała Hellera*. Kraków: Copernicus Center Press, pp.133–153.

Nagel, T., 2012. *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False*. New York: Oxford University Press.

Penrose, R., 2006. *The Road to Reality: A Complete Guide to the Laws of the Universe*. 6th print. New York: Alfred AKnopf.

Perzanowski, J.W. and Sytnik-Czetwertyński, J., 2016. *Jest czyli Rzecz o filozofii bytu*. Toruń: Wydawnictwo Adam Marszałek.

Rovelli, C., 2016. Meaning = Information + Evolution. *arXiv:1611.02420 [physics]* [Online]. Available at: <http://arxiv.org/abs/1611.02420> [visited on 15 December 2021].

Skowron, B., 2020. Virtual objects: Becoming real. *Horizon, Fenomenologiceskie Issledovania* [Online], 9(2). https://doi.org/10.21638/2226-5260-2020-9-2-619-639.

Turek, K., 1978. Filozoficzne aspekty pojęcia informacji. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (1), pp.32–41.

von Weizsäcker, C.F., 1982. *Die Einheit der Natur*. München: Deutscher Taschenbuch Verlag.

Wilczek, F., 2016. *A Beautiful Question: Finding Nature's Deep Design*. London: Penguin Books.

Zurek, W.H., 2002. Decoherence and the Transition from Quantum to Classical — Revisited. *Los Alamos Science* [Online], (27). Ed. by B. Duplantier, J.-M. Raimond and V. Rivasseau, pp.2–25. https://doi.org/10.1007/978-3-7643-7808-0_1.

Machine learning and essentialism

Kristina Šekrst
Sandro Skansi
University of Zagreb, Croatia

Abstract

Machine learning and essentialism have been connected in the past by various researchers, in order to state that the main paradigm in machine learning processes is equivalent to choosing the "essential" attributes for the machine to search for. Our goal in this paper is to show that there are connections between machine learning and essentialism, but only for some kinds of machine learning, and often not including deep learning methods. Similarity-based approaches, more connected to the overall prototype theory, spanning from psychology and linguistics, seem more suited for pattern recognition and complex deep-learning issues, while for classification problems, mostly for unsupervised learning, essentialism seems like the best choice. In order to illustrate the difference better, we will connect both paths to their sources in other disciplines and see how human psychology influences our decision in machine-learning modeling as well. This leads to a philosophically very interesting consequence: even in the setting of supervised machine learning, essences are not present in data, but in targets, which in turn means that the categories which purport to be essences are in fact human-made, and hand-coded in the targets. The success of machine learning, therefore, does not give any substantial evidence for the independent existence of essential

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 171–196

properties. Our stance here is to state that neither the existence nor the lack of "essential" properties in machine learning can lead to metaphysical, i.e., ontological claims.

Keywords

essentialism, machine learning, accidental properties, similarity-based approach, pattern recognition, modal necessity.

Essential and accidental properties: introduction

The purpose of this paper is to show that the existence of essential-like features in machine learning, or the lack of them, cannot provide an ontological commitment.[1] Researchers have connected machine-learning practices with essentialist and anti-essentialist stances, but we feel that such claims ignore that both "essentialist" and "anti-essentialist" paradigms in machine learning are both influenced by human psychology and have no real consequence on the verification of whether there *are* essential properties in nature or not.[2]

The outline of the paper is as follows. First, we will give a brief overview of what philosophical essentialism is and mention the scarce research on (anti-)essentialism in machine learning. Next, we will provide insight into the basics of machine-learning paradigms, namely supervised and unsupervised learning. The notion of essential properties is often connected to supervised learning, but we would like to

[1] The authors would like to thank the anonymous reviewers for their detailed analyses and insights.

[2] It is necessary to distinguish between ontological commitments regarding nature and ontological commitments that are necessary in every AI system (we call this difference an ontological gap). The former are the subject of this article, and the latter were analyzed by Krzanowski & Polak (2022a; 2022b).

find out why, so we will connect it to psychological essentialism and the development of human epistemological stances. We will notice that prototype theory seems closer to human understanding, and it can be seen as present in both supervised and unsupervised paradigms, even though they might be a better fit for the latter. The notion of a *feature vector*, as a collection of properties, will be connected to psychological prototypes. Last, we will observe how in both supervised and unsupervised learning, the human factor involved guides the learning, but this cannot be equated with the existence or non-existence of essentialism. Namely, some supervised tasks are better for some real-life or mathematical problems, while some unsupervised tasks are better for others. The question of essentialist-like or anti-essentialist paradigm here is just a question of using the right tool for your problem, and not an ontological consequence.

Philosophical essentialism

An *essential* property of an object is a property that an object must have, while an *accidental* one is the one the object happens could have, but that it could lack. That is, in modal terms,[3] we are talking about *necessity* and *possibility*,[4] respectively (Robertson Ishii and Atkins,

[3] Standard modal characterizations have been disseminated with the works of Ruth Barcan Marcus and Saul Kripke. Kripke's work on semantics is taking the truth of a formula relative to a possible world, since its truth value depends on what is true in accessible world. See Barcan Marcus (1993) for a modality synthesis and Kripke (1972) for Kripke semantics.

[4] There are, of course, differences between logical and (meta)physical possibilities. Something might not be a logical contradiction, but still be (meta)physically impossible, i.e., not conforming to the laws of nature, for example, a man travelling faster than the speed of light. The exact details of such classifications, especially between physical and metaphysical possibilities, are a matter of debate.

2020): an object is going to possess the essential property in all possible worlds, but for an accidental one, there is a possible world in which an object lacks such a property. *Essentialism* is a standpoint in which (at least) some objects have (at least some) essential properties (Robertson Ishii and Atkins, 2020). For example, an essential property of Socrates is to have originated from his parents but is not essential for him to have brown hair. An essential property of a dog is certainly not brown hair since there are dogs of other colors. In philosophy, some essential properties are not a matter of much debate. For example, a dog had to have some biological origin. *Canis canis* is also a dog. But in order to pursue the matter further, there might be various objections to candidates for essential purposes.[5] Often, a dog is considered a "wolf-like descendant", where various breeds might not conform to this ideal, along with the notion of "having an upturning tail".

The former is the reason why there are various kinds of essentialism in philosophy. Standard Aristotelian essentialism also deals with necessities, and in his categories, he was researching properties that all the members of the category have in common, without which, they cannot be members of that category.[6] One of the most famous criticisms comes from Wittgenstein,[7] who observed the debate from a linguistic angle and stated that words can mean innumerable

[5] A concise description of the debate is provided by Cartwright (1968, p.615): "What are the essential attributes of, say, Dancer's Image? No doubt it will be counted essential that he is a horse and accidental that he was disqualified in this year's Kentucky Derby. But what of the attribute of being male, or of being a thoroughbred, or of not being a Clydesdale stallion? Here, I suppose, essentialists may disagree. Indeed, a reasonable essentialist might well take the position that these are hard cases that admit of no clear decision."

[6] For more details on Aristotelian essentialism, see Aristotle (2014) and Matthews (1990).

[7] See Cohen (1968) for more details.

things depending on their use, paving the way to modern pragmatics. Probably the most common standpoint takes into account that both minimal and maximal essentialism apply. *Maximal essentialism* states that all of any given object's properties are essential to it, while *minimal essentialism* presupposes that there are no limits to the ways a given object might have been different from its current actual state, and the only essential properties seem to be trivial ones, like "being *F*" or "being non-*F*" for any property *F* and "being self-identical" (Robertson Ishii and Atkins, 2020). For the purpose of this paper, we will consider the most common stance as our starting point: maximal and minimal essentialism both hold. The mentioned doctrine that at least some objects have at least some essential properties is the most common one (Robertson Ishii and Atkins, 2020).[8]

Previous work on machine learning and essentialism dealt with various types of machine learning under the same hood, connecting them often to essentialist ideas. Works of Pelillo (2013), Pelillo and Scantamburlo (2013) seem to be the most prominent ones. Tunç (2015) follows Pelillo's (2013) ideas but mostly focuses on epistemology and inductive inference, emphasizing abstracting, idealization, and theoretical variables in machine-learning research. Duin (2015) provides an anti-essentialist approach in pattern-recognition systems, claiming that in most of the applications in pattern recognition, there is no known, small set of essential features (a notion we agree with). Our goal is to show how various cases of essentialist-like and non-essentialist-like stances can be seen manifested in machine-learning choices, but that does not mean we are talking about real essentialist or anti-essentialist ontology.

[8] Explicitly stated by Mackie (2006). For more details about various types of essentialism, see (Robertson Ishii and Atkins, 2020).

Machine-learning basics

Machine learning, as a part of artificial intelligence and computer science, is a field of approaches and methods that use data in order to improve their performance on some problems. Artificial intelligence can be seen as a certain type of philosophical engineering (Skansi, 2018, p.vii): we want "to build machines that can think, [. . .] understand the meaning, act rationally, cope with uncertainty, [. . .] handle and talk about objects". In a nutshell, we are replicating standard philosophical concepts. It is no wonder that philosophical concepts are deeply embedded in their methods as well but may seem hidden underneath technical layers.

In machine learning, data is usually split into *training* and *test* data, the same way a student learns methods and approaches to some problems and gets previously unseen ones in an exam. Such an approach, compared to learning in the presence of a supervisor or a teacher, is called *supervised learning*: an algorithm learns from *labeled* data and is able to predict outcomes on previously unseen data. For example, if we had a dataset consisting of various pictures of animals, and we wanted to train the algorithm to recognize cats, we would want it to be able to somehow point out what is *essential* for an animal to be classified as a cat. An important part of supervised learning is therefore the act of *classification*: a certain object of interest possesses or does not possess certain property, i.e., it is or is not a member of a class. A certain image of a dog might be marked as 98.6% dog if it is very close to all of the properties that seem to be essential for classifying a picture of an animal as a dog. However, a cat might have some properties, such as four legs and a tail, but that would be a low percentage. In another class of problems, there are *regression* problems, in which the algorithm is predicting continuous

values. For example, given previous real-estate prices in a certain area, predict the prices for the next couple of years. Here, we would be dealing with real numbers instead of binary Boolean classifications.[9]

To summarize, a supervised machine-learning algorithm receives a set of training data points (a point in space where the axes are the properties given) and labels (row vectors), and in this phase, the algorithm creates a hyperplane—a decision boundary that helps classify its data points—by adjusting its internal parameters (Skansi, 2018, pp.55–56). This phase is the training phase that receives inputs as row vectors with corresponding labels (called training samples). In the next, predicting phase, the trained algorithm takes a number of row vectors, but this time without labels and creates the labels with the hyperplane (Skansi, 2018, p.56). In other words, "the learner receives a set of labeled examples as training data and makes predictions for all unseen points, [. . . a scenario commonly] associated with classification, regression, and ranking [i.e. ordering items to some criterion] problems" (Mohri, Rostamizadeh and Talwalkar, 2018, p.6).

Another type of machine learning, *unsupervised learning*, handles various datasets without any explicit instructions or labels. That is, the learner receives unlabeled training data and makes predictions for all unseen points, and "since, in general, no labeled example is available [. . .], it can be difficult to quantitatively evaluate the performance of a learner" (Mohri, Rostamizadeh and Talwalkar, 2018, p.7). Unsupervised learning encompasses a broad definition of learning without labels or targets, but this broad definition begs the cognitive question of how we learn without feedback (Skansi, 2018, p.70). In the previous case, in order for the computer to learn what is a dog,

[9] As a side note, most of the algorithms do not predict using Boolean outcomes such as 0 or 1 for not being or being a dog, but as a matter of a percentage. In such cases, we are effectively talking about fuzzy intervals.

we had to correctly label dogs or provide a list of properties in other supervised examples. Here, a computer is seemingly on its own: for example, neural networks[10] tend to automatically find structures in the data by analyzing useful features. Data is often grouped into *clusters*, and then it is easy to see the outliers, anomalies (for example, for fraud detection), associations (for recommender systems), and similar connections.

We might start to notice something interesting here. First, if we are telling the computer while we are labeling the data that something is or is not a certain kind of object, we are effectively taking a certain essentialist stance. Intuitively, there seems to be something *essential* in all of the properties that make a cat a *cat*. In various cases of supervised learning, we might list a number of features that we could consider important. For example, my algorithm might be tracking pointy ears, four legs, two eyes, and fur. But such an algorithm might recognize dogs and rabbits as well but miss some dogs without pointy ears. And we are not even starting to talk about three-legged dogs and similar "obviously" accidental properties. Second, it all boils down to starting human decisions. This seems like a trivial claim, and from a description of supervised learning, it is rather intuitive. Blaming it on the data might seem like a common excuse in machine learning, but recently, AI ethics has dwelled on the questions of initial data handling and responsibilities.[11] However, why did we choose some features on top of others? The answer might lie in human psychology.

[10] See Skansi (2018) for an introduction to deep learning.

[11] A famous example of an accidental algorithmic breach of ethics includes machine-learning racism in tagging black people as gorillas. See Zhang (2015).

Psychological essentialism

Gelman (2004, p.405) describes how once children learn a new fact about one member of a category, they generalize the fact to other members of that category, even if they look substantially different. By four years of age, children display subtlety and flexibility when they make category-based inductive inferences. For Gelman (2004, p.405), properties seem to be "fixed at birth", demonstrated by the following experiment. A child might learn about a newborn kangaroo that was switched at birth, and then went to live with goats. The child was then asked whether the animal would be good at hopping or climbing, or if would it have a pouch or not. Turns out, preschool children typically reported that it would have been good at hopping and have a pouch, something that seems *inherent* to kangaroos even for children. Such an understanding seems to appear by about six years of age, and it might be as early as four years of age: the time when children reason about animals, plants, and social categories (Gelman, 2004, p.406).

By the age of two, children view *causes* as vital to what something is (Gelman, 2004, p.406). This is interesting from a philosophical standpoint. *Causal essentialists* hold that a property essentially bears its causal and nomic relations (Gibbs, 2018, p.2332). Such a stance constrains what is possible and rules out possibilities where a property bears causal and nomic relations differently from how it actually bears them (Gibbs, 2018, p.2334). It seems that the notion of a cause and similar notions of origins seems to be closely tied to our early-childhood understanding of such relationships. There are some intriguing mistakes here: Gelman (2004, p.406) mentions that children sometimes can be more "nativist" than adults. For example, five-year-olds claim that a child switched at birth will speak the language of their birth parents rather than adoptive ones. We know that this is

not the case, but it is intriguing to see how an essentialist "feeling" might not always be correct if we take cognitive development as our guideline. Causality is central to children's categories, claims Gelman (2004, p.406), since it provides consistent domain-specific causal explanations for the properties that members of a category share. That is, category membership is stable over transformations (a dog cannot be transformed into a cat), and internal properties seem to be salient to young children. In a way, this is how computers behave as well: learning from observations and from their parents and other people. In the case of supervised machine learning, that is a combination of a prelabeled dataset and learning from data.

Here, the notion of a *feature* comes in handy as an individual measurable property. In character recognition, features might be shapes and pixels, and in voice recognition, frequency, noise, and strength. In computer vision, we might be talking about blobs, i.e., regions in images that differ in properties from the rest of the surrounding regions, for example, in color or brightness. Basically, it is a collection of information used for future problem resolution. In the case of classification, this may be compared to children learning about classes, memberships, and categories. But, how to describe a dog, say, using words? What are some essential properties the algorithm would be searching for? For example, a type of face, number of various body parts, color, etc. Children learn to recognize various members of the class and then generalize and use this knowledge in novel situations, i.e., previously unseen examples of that class. We want the computer to follow a similar process. In order to generalize well, a good selection of features needs to be selected. In the next section, we will observe how such a process is followed in machine learning and how the question of feature selection has important philosophical consequences.

Features and prototypes

As mentioned, *features* tend to be measurable properties that are successful in discriminating and differentiating between different categories of data. For example, in face detection (Bishop, 2006, p.3), we aim to find features that are not only fast to compute but also preserve useful discriminatory information enabling faces to be distinguished from non-faces. The study of feature selection finds its practical needs in machine learning, where a learning algorithm constructs a description of a function from a set of input/output instances through interaction with the world. Machine learning is more concerned with non-continuous features, while pattern recognition deals with continuous ones. It is not the same, for example, to classify something as a dog or not, compared to finding a face or another pattern or blob inside an image (Liu and Motoda, 1998, p.2). Liu and Motoda (1998, p.2) state that many forms of representations for machine-learning functions are available, including first-order logic, which is interesting from a philosophical standpoint, or weighted networks, but they have focused on features since they are 1) *primitive* 2) *convenient* 3) *independent* 4) *widely used* 5) *reasonably general*, i.e., powerful for many applications.

The first condition is the most important one for a metaphysics approach, and they define it as "the basic units for defining a problem, a domain, or a world to be observed, and do not require much effort from human experts to design them". Taken into account that feature selection tasks often fall into the hands of non-metaphysicists, there is a hunch of an innate human ability to generalize and select something that might, at least in the layman's sense of the word, seem essential for the object in question. Features are also called *attributes*, *properties*, or *characteristics* and can be discrete, continuous, or complex

(Liu and Motoda, 1998, p.3). For example, a dataset consisting of various hairstyles might have a feature of [color], which would take color names or RGB codes as its discrete value, [hair_length] may be a continuous numerical value in centimeters or inches, while there might be a Boolean [is_dyed] with true or false discrete values.

Trying to describe a certain object by finding whether it has or has not some constitutional properties, along with describing them, is not a novelty of machine learning. The same formal approach was popular with the advent of structural linguistics. Since phonology studies its basic units—phonemes, morphology analyzes morphemes, and syntax inspects sentence elements such as subjects, objects, and phrases, it was natural to try to find a basic unit of meaning that would make semantics an equal member of the formalized grammatical discipline ensemble.

Semic analysis was the first approach that aimed to find minimal units of meaning, which later developed into *componential analysis* within the standard structuralist framework. In particular, Pottier (1964) analyzed various types of chairs in order to find out what are the minimal features needed in order to distinguish between them. For example, they might have a back side or not, might have arms or not, can be fixed or folding, can have one seat or several seats, etc. One can see that we are already dealing with both discrete and continuous values here. A classic example in the componential analysis is how to describe various words for human beings in various stages of their lives, taking into account their gender. A *man* can be described as [-woman] and [+adult] or [+man] and [+adult]. Here we are dealing with Boolean man/female and adult/not adult, which does not reflect the fuzzy values of such categories, but structuralist linguistics was extremely focused on binary oppositions. Next, a *woman* would be [+woman] and [+adult] or [-man] and [+adult], a *girl* would be

[+woman] but [-adult] or [-man] and [-adult], while a *boy* could be described as [-woman] and [-adult] or [+man] and [-adult].[12] Such an approach has been developed and changed but is still used in semantics, which, as one of its tasks, analyzes the internal structure of a word by finding distinct and minimal components of meaning (Palmer, 1981, p.108). In such a framework, we might differentiate our *dog* from a *wolf* by finding distinct features. Both are certainly [+animal] and [+canine], but we might add [+domesticated] to the dog and [-domesticated] to the wolf. Such choices often seem arbitrary and there is no consensus on what the most basic properties of objects or classes of objects are, and it would also seem necessary to connect not only machine-learning feature selection with philosophy but linguistics and psychology as well.

We have mentioned that a strict binarist approach may seem inadequate in many cases. Departing from a standard Aristotelian notion of fixed categories, Eleanor Rosch (1973) introduced the *prototype theory* in which there is a graded degree of belonging to a certain category: some members are more central than others. For example, whatever essential properties of a bird might be, it seems somehow intuitive that in this—perhaps arbitrary—category there are some *more prototypical members* or examples than others: a sparrow is a more prototypical bird than an ostrich or a penguin. But this seems culturally anchored in both time and space, an apple is a more prototypical fruit in Europe, but other fruits might be better examples in Africa, such as bananas.

In machine learning, a feature does not have to be a binary Boolean, it can also be seen and created as a certain prototype. For ex-

[12] Such a method was formed on the basis of Prague structuralist school dealing with phonology. A phoneme has a set of discrete properties, for example *b* would be [+voiced], while *p* would be [-voiced], but both would be [+labial] plosives.

ample, in image recognition, there is a need to give the best prototype for a category. In the case of supervised learning, if we are training our models to recognize birds, and we are only using edge-case birds, we are not using the most generalized and best prototype or a versatile dataset consisting of central and edge-case members. The majority of images presented in a labeled training dataset would be close to being a prototype of the category. If we wanted to recognize apples, a rotten or a half-eaten apple would not be a prototype but would be a wanted member of the class, and if we wanted to recognize cats, a one-eyed cat without ears would not be a prototypical image, but we would somehow like to get the essentials with our prototypes in order to also include this as a result. In this case, we would expect percentages stating the probability that something is a dog or a cat to be higher for prototypical members, possessing all the necessary features, and less for edge-case or less prototypical members of a category.

Essentialist paradigm(s) in machine learning

It seems intuitive and obvious that supervised machine learning incorporates some kind of essentialism. That is, we are either given discrete or continuous features in datasets that are used for our predictions, usually whether something is a member of a class or not. But there are other kinds of machine learning, and we must not ignore the notion of unsupervised learning. We have already mentioned unsupervised learning, in which a model tries to establish regularities, clusters, or patterns in previously unseen data. This can be compared to the process of human learning at an early age, in which a human being tries to generalize the already acquired knowledge. Consider this, even if you are getting an unlabeled dataset of weird alien creatures, you will

most certainly be able to connect similar ones together in groups or do classifications, even if you do not know what *is* actually in the background of your dataset. We would like the computer to do the same. For example, if we trained our models on a certain map, they might recognize landmasses, developed areas, forests, or wetlands and group them together, by finding similarities between them. In non-visual data, you might be presented with some numbers, say bank transfers, and you might connect the usual activity into groups, while the outliers might be suspicious.

Pelillo and Scantamburlo (2013) were one of the pioneers of trying to connect machine learning with metaphysics. For them, the majority of traditional machine learning techniques are centered around the notion of a "feature", which we have observed. However, they note that there are numerous application domains where either it is not possible to find satisfactory features, or they are inefficient for learning purposes. Such examples might include cases when experts cannot define features in a straightforward way (e.g., protein descriptors vs. alignments), cases when data are highly dimensional (e.g., images), situations when features consist of both numerical and categorical variables (e.g., person data, like weight, sex, eye color, etc.), or in the presence of missing or inhomogeneous data.

In his overview of pattern recognition, which is mostly unsupervised, Pelillo (2013) states that features are *essential* properties. He reports Watanabe (1985) stating that "under all works of pattern recognition lies tacitly the Aristotelian view that the world consists of a discrete number of self-identical objects provided with, other than fleeting accidental properties, a number of fixed or very slowly changing attributes. Some of these attributes, which may be called 'features', determine the class to which the object belongs. Pellilo (2013, p.2) reaffirms that the goal of a pattern recognition algorithm

is to discern "the essences of a category" and that we should talk about an essentialist paradigm in machine learning. We have already mentioned Rosch's (1973) work on prototypes, which Pelillo (2013, p.2) uses to illustrate the "multifaceted nature of real-world categories" and emphasizes that for anti-essentialist stances, relations are in focus. That, of course, does not have to be the case, the main idea for anti-essentialism is to claim accidentality: there are possible worlds in which the object has the property in question and possible worlds in which it does not. But he does emphasize that the feature-based aspect is a *reductionist* position since objects are seen in isolation and overlook relational or contextual information (Pelillo, 2013, p.1).

The notion of a *feature vector* is often used in machine learning: an n-dimensional vector that serves a purpose of a collection of features. For example, just as a red/green/blue combination will form a single color, a certain combination of features will be used in machine-learning tasks to better identify objects or predict values. Pelillo (2013, p.3) emphasizes that the community has focused on feature-vector representations, rather than on single, standalone features. In computer vision and pattern recognition, each object is described in terms of a vector of numerical attributes and mapped to a point in a Euclidean vector space, so that the distances between the points reflect the similarities and dissimilarities between the respective objects (Pelillo, 2013, p.3). Pelillo emphasizes the recent trend in *similarity-based techniques*, which are still not challenging the traditional paradigm but work with graphs or structural representations to find objects or values that seem to be closer according to some criterion (Pelillo, 2013, p.4). We have to note that such an approach is analogous to a prototypical relationship, where members are grouped around a prototype in a certain graph-like manner. A green apple is

more similar to the prototype of a red apple than a red strawberry, and if such connections would be shown as a weighted graph, then we would expect a less expensive traversal to a red apple.

Accidental properties in machine learning

The processes and disciplines of *feature selection* and *extraction* show us that there is a strong presupposition that something as essential as a feature exists. There is no doubt that machine learning today is still enveloped in a strong essentialist paradigm. In feature engineering,[13] a system automatically discovers representations needed for feature detection. For example, it finds close points (neighbors) in a graph and clusters data around, say, percentages. If feature engineering is an essentialist stance, what kind of essentialism is it? It seems that is not maximal, but also not minimal, we would expect it to lie somewhere in between, judging by its success factor.

Here, what is interesting is that, unlike in human-led feature selection, automated feature engineering may use features that a philosopher would deem completely accidental, but it would still do a great job in classification or similar predictions. That is, deep-learning feature engineering does not have to correspond to some natural kinds or essential properties: it is not really essentialism, but a certain kind of accidentalism.

Namely, sometimes, features even outside deep learning that generate best models are often surprising and maybe even lucky correlations.[14] A famous example is a system (Lapuschkin et al.,

[13] For more details about feature engineering, see Zheng and Casari (2018).

[14] Some would argue that such processes might fall under the umbrella of *unexplainable AI*, if we are dealing with multiple layers within deep neural networks, but in

2019) performing horse recognition that learned to cheat by looking for the copyright watermark in horse images instead of finding some horse-essential features.[15]

When it is led by humans, that does not mean that there is an omniscient metaphysicist in computer engineers deciding what is essential and what is not. There are two important problems in machine learning. The first one is *underfitting*, the case in which a model is too general and does not fit the data property. For example, if we were doing dog recognition, from the training set, our underfitted model would consider that necessary features would be to have pointy ears and tails. In this case, we might recognize cats and rabbits too. An *overfitted* model has the opposite problem, it too closely responds to training data, and it is too specific. Basically, as if you only knew how to solve problems that appeared in your homework, but you are unable to solve the same problem when the numbers are replaced with other numbers. Our model might only recognize white and fluffy dogs with grey spots on their backs. Such a case might also be a result of bad feature engineering in the first place. Using automated feature engineering actually reduces the overfitting of your models, taking into account the standardized method of figuring out which one of your selected features might cause problems for your model to be too specific. We might imagine a case in which that also might seem like

---

the worst cases, "unexplainable" is not *impossible* to test or retrace, just *not easy*. We deem that the problem is not in unexplainability, but usually in the human inability to comprehend the data or the wrong (perhaps "accidental") approach taken.

[15] There are various legends and "folk tales" stating variations of a tank story, in which Russian tanks were photographed during daytime, unlike British tanks, so the AI system used that to its advantage. Most of such stories are farfetched but they do serve a purpose of illustrating a *possible* way an AI system might come to the right conclusion using the wrong method.

an essential property, but not for machine-learning purposes. Properties chosen or discovered might not be relevant or essential but make the model perform well.

Machine learning or essentialism?

Our previous conclusion might imply two separate things. Either there is an anti-cybernetic stance in which human learning that encompasses a certain kind of innate essentialist knowledge is a different process in machine learning, or that, for practical purposes, knowledge of essential properties is not a necessary prerequisite for everyday classifications and predictions. The latter seems more intuitive. It does seem that a similarity-based approach, mimicking the prototypical relationships found in psychological and linguistic research, may work well in various human and machine usages, along with a combination of properties (features) together with their relations (cf. feature vectors). For some machine-learning tasks, pure essentialism, often a binary or Boolean one, works best. We believe that essentialism and anti-essentialisms are not binary choices a computer scientist or a philosopher must make in order to describe how processes are being generated and run in machine learning paradigms today, but it is a matter of choice *for a specific type of task*. There is no essentialism equated with machine learning, but there is both essentialism and anti-essentialism for specific tasks. For some classification tasks and simple pattern recognitions, essentialist features are often the best choice, and for others, systems will work better with combinations of these properties. For unsupervised learning and pattern recognition, prototypical systems, i.e., similarity-based approaches, perform better. A philosophical take here is that, at least in machine learning, there is

no ontological obligation towards either of these stances, but rightful usage for rightful tasks. The choice of your machine-learning system, and therefore, a supervised or unsupervised approach, will depend on the type of task in question: *what performs better*. It is just a matter of technical performance that has no metaphysical consequences of the existence of essentialism or anti-essentialism.

From a psychological standpoint, Gelman (2005) has shown that essentialism is present in our everyday choices and is a reasoning heuristic readily available to both children and adults. As human beings, we seem to be hard-wired to search for parts and underlying structures. She claims that preschool children and adults from a variety of cultural contexts expect members of a category to be alike in a non-obvious way. That is, we treat "certain categories as having inductive potential, an innate basis, stable category membership, and sharp boundaries" (Gelman, 2005). It is no wonder that essentialist research has emerged as a metaphysical position. However, often, in our everyday practice, we are proven wrong, and that goes for our early childhood as well: Gelman's (2004) example of children being more nativist than adults. If essentialism might not always be the right choice for humans in various contexts, then the characterization of machine learning as an "essentialist" paradigm only reflects our inner psychological phenomena.

In philosophy, such an idea is present in the stance of conventionalism. Conventionalism seeks to expose conventions likely to be mistaken for truths (Ben-Menahem, 2006, p.2). This relativistic view is close to our claim that both supervised and unsupervised learning are plagued with human psychological categories that do not say anything about the possibility of objective categories, but only that we might or might not interpret conventions in various ways, even in essentialist and anti-essentialist terms.

As we have shown, the dichotomy should have never been the one about the differences in learning by humans or machines since these epistemic differences do not exist. The first reason is simple: machine learning is modeled after human learning, and only after the initial modeling is fine-tuned to make it computationally feasible. It is "essentially" the same by design. The differences are, again by design, accidental and purely due to different hardware/wetware. The second reason is more cybernetic in nature: if we are to develop a learning theory, it should be able to be as general as possible. Today one would never accept a psychological theory that only explains fear in adults or anxiety in women. Even though we might need to limit our theory in such a manner until further research is conducted, we would never accept this to be a completed theory. A theory of learning which would explain learning in children but not adults would likewise be incomplete and unacceptable except as a work in progress. This theory would be expanded to adults, people with disabilities, and to different cultures. After all, this is supposed to be a general theory of learning. Even though limiting the theory to humans might sound appealing, one could speculate that there will be more than a handful of researchers interested to see how such a theory applies to apes or dogs. Xenobiologists might take an interest too, as could AI researchers. Social scientists and cultural anthropologists might be also tempted to see if such a theory can describe models of societal learning or cultural integration. The point here is that the cybernetic call is a very natural force in scientific expansion and research, one that is to be expected, and one we had seen in a number of fields, perhaps the most recent and interesting one being social physics (as a branch of social network analysis). The insights gained in this fashion not only have huge practical benefits, but they do tend

to encompass a basic scientific curiosity, which no philosophy of science can avoid: "they say X and Y are not connected, but what happens if I use X on Y?".

The true dichotomy still present is a wholly different one. In fact, it is the same one that René Descartes described half a millennium ago: do the categories present in my mind have objective validity?[16] The easiest way to a positive answer is essentialism, which claims that the categories in our minds are formed via the essential properties present in the world. And machine learning, a new paradigm where machines are finally intelligent enough, is believed by many to show exactly this. If machines can learn the same things we do, then obviously the categories used are not intrinsically human. If machines can learn this by crunching data obtained from the world, then the categories are in fact present in that very data as essential properties. Machine learning is, on this account, simply a family of algorithms capable of extracting not just information from data, but essential properties as well. As we have shown, this view is wrong, since: (i) this could in theory hold true only for supervised learning, and more importantly (ii) supervised learning is defined via its use of targets or labels which are *man-made*. Since they are man-made, this means that human annotators bring in their categories "cat/dog", "animal/non-animal", "happy/sad", etc., and connect this to actual data, e.g., pixel values, or numeric data. The machine-learning algorithm then extracts this connection and applies it to previously unseen data. But the essential properties are not the ones discovered by the algorithm, they are brought in by human annotators, and do not have to reflect the "real" ontology at all. Even in the case of unsupervised learning, the features are being clustered and interpreted by humans, bringing again their own categories into play.

[16] See Descartes (1641; English translation: 1991) for more details.

Essentialist and anti-essentialist stances are both present in supervised and unsupervised learning, but we have pinpointed a couple of claims. First, supervised learning is easily connected with essentialism, but we wanted to pinpoint that it does not bear an *ontological commitment* to the existence of such features. Even though the view itself that humans are creators of essential features in machine learning might seem trivial, it does not say anything about ontology, but it says a lot about human psychology. Second, we might talk about the anti-essentialist stance in unsupervised learning (as Duin (2015) does), but this again is a strong ontological claim. Our goal was to show that unsupervised-learning approaches follow the prototypical learning and categorization model, inherent to human psychology, which also might be something the model creators are bringing to the model itself. The choice of supervised or unsupervised methods, which some might equate with essentialist or essentialist stances, actually does not exist since the choice depends on the problem we want to solve. Machine-learning systems do not discover anything about background ontology, but they do show us human epistemology and psychology present in seemingly competitive stances.

Bibliography

Aristotle, 2014. Categories. In: Jonathan Barnes, ed. *The Complete Works of Aristotle: The Revised Oxford Translation, One-Volume Digital Edition*. 6. print., with corr. Vol. 71:2, *Bollingen series*. Princeton, N.J: Princeton University Press, pp.25–70.

Ben-Menahem, Y., 2006. *Conventionalism : From Poincare to Quine* [Online]. Cambridge; New York: Cambridge University Press. Available at: <https://search.ebscohost.com/login.aspx?direct=true&db=e000xww&AN=529339&lang=pl&site=ehost-live> [visited on 13 January 2023].

Bishop, C.M., 2006. *Pattern Recognition and Machine Learning*, *Information science and statistics*. New York: Springer.

Cartwright, R.L., 1968. Some Remarks on Essentialism. *The Journal of Philosophy* [Online], 65(20), pp.615–626. https://doi.org/10.2307/2024315.

Cohen, M.F., 1968. Wittgenstein's anti-essentialism. *Australasian Journal of Philosophy* [Online], 46(3), pp.210–224. https://doi.org/10.1080/00048406812341181.

Descartes, R., 1641. *Renati Des-Cartes Meditationes de prima philosophia, in qua Dei existentia et animae immortalitas demonstratur.* [Online]. Paris: Michael Soly. Available at: <https://gallica.bnf.fr/ark:/12148/btv1b86002964> [visited on 25 August 2021].

Descartes, R., 1991. Meditations on First Philosophy. *The Philosophical Writings of Descartes, vol. 2* (J. Cottingham, R. Stoothoff and D. Murdoch, Trans.). Cambridge: Cambridge University Press, pp.1–63.

Duin, R.P., 2015. The dissimilarity representation for finding universals from particulars by an anti-essentialist approach. *Pattern Recognition Letters* [Online], 64(C), pp.37–43. https://doi.org/10.1016/j.patrec.2015.04.015.

Gelman, S., 2004. Psychological essentialism in children. *Trends in Cognitive Sciences* [Online], 8(9), pp.404–409. https://doi.org/10.1016/j.tics.2004.07.001.

Gelman, S.A., 2005. *Essentialism in Everyday Thought*. Available at: <https://www.apa.org/science/about/psa/2005/05/gelman> [visited on 12 January 2023].

Gibbs, C., 2018. Causal essentialism and the identity of indiscernibles. *Philosophical Studies* [Online], 175(9), pp.2331–2351. https://doi.org/10.1007/s11098-017-0961-y.

Kripke, S.A., 1972. Naming and Necessity. In: D. Davidson and G. Harman, eds. *Semantics of Natural Language* [Online], *Synthese Library*. Dordrecht: Springer Netherlands, pp.253–355. https://doi.org/10.1007/978-94-010-2557-7_9.

Krzanowski, R. and Polak, P., 2022a. Ontology and AI Paradigms. *Proceedings* [Online], 81(1), p.119. https://doi.org/10.3390/proceedings2022081119.

Krzanowski, R. and Polak, P., 2022b. The Meta-Ontology of AI systems with Human-Level Intelligence. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (73), pp.197–230.

Lapuschkin, S. et al., 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications* [Online], 10(1), p.1096. https://doi.org/10.1038/s41467-019-08987-4.

Liu, H. and Motoda, H., 1998. *Feature Selection for Knowledge Discovery and Data Mining*. Norwell, MA: Kluwer Academic Publishers.

Mackie, P., 2006. *How Things Might Have Been: Individuals, Kinds, and Essential Properties* [Online]. 1st ed. Oxford: Oxford University Press. https://doi.org/10.1093/0199272204.001.0001.

Marcus, R.B., 1993. *Modalities: Philosophical Essays* [Online]. New York: Oxford University Press. Available at: <http://catdir.loc.gov/catdir/enhancements/fy0638/91048105-t.html> [visited on 12 January 2023].

Matthews, G.B., 1990. Aristotelian Essentialism. *Philosophy and Phenomenological Research* [Online], 50, pp.251–262. https://doi.org/10.2307/2108042.

Mohri, M., Rostamizadeh, A. and Talwalkar, A., 2018. *Foundations of Machine Learning* [Online]. 2nd ed., *Adaptive computation and machine learning*. Cambridge, MA: The MIT Press. Available at: <https://cs.nyu.edu/~mohri/mlbook/> [visited on 13 January 2023].

Palmer, F.R., 1981. *Semantics* [Online]. 2nd ed. Cambridge: Cambridge University Press. Available at: <http://archive.org/details/semantics00palm> [visited on 13 January 2023].

Pelillo, M., 2013. Introduction: The SIMBAD Project. In: M. Pelillo, ed. *Similarity-Based Pattern Analysis and Recognition* [Online], *Advances in Computer Vision and Pattern Recognition*. London; Heidelberg; New York; Dordrecht: Springer, pp.1–10. https://doi.org/10.1007/978-1-4471-5628-4_1.

Pelillo, M. and Scantamburlo, T., 2013. How Mature Is the Field of Machine Learning? In: D. Hutchison et al., eds. *AI*IA 2013: Advances in Artificial Intelligence* [Online]. Vol. 8249. Cham: Springer International Publishing, pp.121–132. https://doi.org/10.1007/978-3-319-03524-6_11.

Pottier, B., 1964. *Vers une sémantique moderne*. Strasbourg: Klincksieck.

Robertson Ishii, T. and Atkins, P., 2020. Essential vs. Accidental Properties. In: E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy* [Online]. Winter 2020. Stanford, CA: Metaphysics Research Lab, Stanford University. Available at: <https://plato.stanford.edu/archives/win2020/entries/essential-accidental/>.

Rosch, E.H., 1973. Natural categories. *Cognitive Psychology* [Online], 4(3), pp.328–350. https://doi.org/10.1016/0010-0285(73)90017-0.

Skansi, S., 2018. *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence* [Online], *Undergraduate Topics in Computer Science*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-73004-2.

Tunç, B., 2015. Semantics of object representation in machine learning. *Pattern Recognition Letters* [Online], 64(15), pp.30–36. https://doi.org/10.1016/j.patrec.2015.03.016.

Watanabe, S., 1985. *Pattern Recognition: Human and Mechanical*. New York: John Wiley & Sons, Inc.

Zhang, M., 2015. *Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software*. Available at: <https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/> [visited on 13 January 2023].

Zheng, A. and Casari, A., 2018. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. 1st ed. Beijing; Boston; Farnham; Sebastopol; Tokyo: O'Reilly.

The meta-ontology of AI systems with human-level intelligence

Roman Krzanowski
Pawel Polak
Pontifical University of John Paul II in Kraków, Poland

Abstract

In this paper, we examine the meta-ontology of AI systems with human-level intelligence, with us denoting such AI systems as AI_E. Meta-ontology in philosophy is a discourse centered on ontology, ontological commitment, and the truth condition of ontological theories. We therefore discuss how meta-ontology is conceptualized for AI_E systems. We posit that the meta-ontology of AI_E systems is not concerned with computational representations of reality in the form of structures, data constructs, or computational concepts, while the ontological commitment of AI_E systems is directed toward what exists in the outside world. Furthermore, the truth condition of the ontology (which is meta-ontological assumption) of AI_E systems does not require consistency with closed conceptual schema or ontological theories but rather with reality, or in other words, "what is the world" (Smith, 2019, p.57). In addition, the truth condition of AI_E systems is verified through operational success rather than by coherence with theories. This work builds on ontological postulates about AI systems that were formulated by Brian Cantwell Smith (2019).

Keywords

human-level intelligence AI, meta-ontology of AI Paradigm, AI

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 197–230

Paradigm, ontology of AI Paradigm, ontological commitment of AI Paradigm, John Haugeland, Brian Cantwell Smith, Marvin Minsky, Hubert Dreyfus.

Introduction

Artificial intelligence systems have developed over the past 60 years, bringing new solutions to a huge number of practical problems, and they continue to find many surprising and fascinating applications. However, the main goal of AI, namely the creation of a human-like intelligence,[1] is still proving unattainable. Indeed, the AI systems we currently design and implement cannot replicate human intelligence and a human agent's ability to cope with reality (see, e.g., Brooks, 1991; Minsky, 1991; Dreyfus, 2016; Mitchell, 2019; Bołtuć, 2020; Roitblat, 2020; Wooldridge, 2021).

One of the reasons for this failing (in their ability to cope with reality) is, it seems, related to how these AI systems lack a proper ontology or representation of the real world. (For more about the failings of current AI conceptualizations, see, for example, the works of Brooks (1991), Dreyfus (2016), and Smith (2019)).[2] Smith (2019, p.44) proposed four features that AI systems should possess if they are to mimic humans' ability to cope with the real world (i.e., have human-level intelligence). These systems, which we here refer to as AI_E systems, should be embodied, embedded, extended, and enactive

[1] See ft. 3 on AGI for an explanation of human-like intelligence.

[2] To get a sense of the ontologies (and meta-ontologies) of the real world in biological agents, consult Ed Yong's book *An Immense World* (Yong, 2022). Interesting analysis of deep learning systems ontological commitments see (Šekrst and Skansi, 2022).

(see also Käufer and Chemero, 2021). While these features are not ontological per se, but they do imply a commitment to some ontology. What the kind of ontology these four features would entail is a meta-ontological question that we explore in this paper, thus furthering the ideas studied by Krzanowski and Polak (2022).

This paper is organized as follows: First, we define some basic concepts related to the meta-ontological discourse, namely ontology, specifically ontology of computing and AI systems, meta-ontology, ontological commitment, and truth conditions. As these concepts have many interpretations, we need precise definitions to ensure that the subsequent discussion will be understood as intended. Next, we explain the main postulates of meta-ontology for AI_E systems, the topic of this work. In the conclusion, we discuss the inherent limitations of ontologies (which is a meta-ontological problem) for artificial systems such as AI_E, their inability to match human intelligence, and the potential prospects for AI_E. (For more on the problems of ontology in AI systems see, for example, Haugeland (1985) and Fjelland (2020)).

Three things should be borne in mind while reading this paper. First, AI systems with human-level intelligence are often referred to as AGI systems, but with many interpretations for this concept, we avoid using this term to prevent us from drifting into the debate about AGI.[3] Second, this is a study of the meta-ontology of specific AI systems, i.e., AI_E, meaning that the focus of the study is ontology of these

[3] See various references for different conceptualizations of AGI (e.g., Mitchell, 2019, p.40; Fjelland, 2020); general purpose, human-level intelligence (Marcus, 2022); the generic ability of a machine to consciously perform any task that a human can (Swar, Khoriba and Belal, 2022); the intelligence of a machine that is capable of understanding the world (Skuza, 2020); the representation of generalized human cognitive abilities (Lutkevich, 2022); a general-purpose capability, including the ability to broadly generalize to fundamentally new areas (Cassimatis, Bello and Langley, 2008); and the capacity of an engineered system to display the same rough sort of general intelligence as humans (Goertzel, 2015). Creating human-level intelligence

AI systems rather than philosophical ontology. Philosophy forms the background of this discussion, but it is not its main objective. Third, we are not concerned with particular implementations of AI_E systems, which is why we instead study the AI_E system paradigm, which is the all-encompassing conceptual framework of AI_E systems that supports multiple implementations. Thus, when we talk about an AI_E system, we are referring to an AI_E system paradigm rather than a specific implementation. The concept of AI paradigm and its role in this study are explained later in the paper.

Key grounding ideas

The ontology of AI

Ontology can be thought as an empty buzzword or a specific conceptual construct,[4] so we need to position the ontology of AI_E within the world of ontological theories and demonstrate, what it means in this discussion, how it relates to "other" ontologies in philosophy, computer science, and AI. Ontology in computer science and AI systems (e.g., Sánchez, Cavero and Martínez, 2007; as well Guarino and Gia-

---

was always the aim of AI research, as attested to by Yann LeCun's recent claim "Getting machines to behave like humans and animals has been the quest of my life" (reported in 2022 MIT Technology Review (Heikkilä and Heaven, 2022)).

[4] "We must be careful in reading [auth. any] philosophical works on ontology, when the author speaks of 'ontology' without qualifications, not to confuse the intended sense of the world with any of the alternatives" (Jacquette, 2002, p.3). There is also a confusion between ontology and metaphysics. Some authors see ontology as the ultimate study of reality (e.g., Jacquette, 2002; Stróżewski, 2004; or Perzanowski, 2015), and metaphysics as being "after physics", some others see ontology as a part of metaphysics (see e.g., Van Inwagen, 2009). In AI literature because ontology takes on a very concrete garb (of an engineering domain) metaphysics is a rare term so the confusion is not so visible.

retta, 1995; Guarino, Oberle and Staab, 2009; Swar, Khoriba and Belal, 2022) takes on different meanings to that of philosophy (e.g., Jacquette, 2002; Stróżewski, 2004; Baker, 2007; Chalmers, Wasserman and Manley, 2009; Effingham, 2013; Berto and Plebani, 2015; Perzanowski, 2015; Thomasson, 2015; Hofweber, 2021).[5] The philosophical concepts of ontology, however, are fundamental to those used in specific applications (see the comments of Jacquette, 2002, p.XII). This is therefore where we begin.

In philosophy, ontology is the study of being as it is (i.e., "what is"), so it is about "being" in the most general sense.[6] More specifically, in its purely philosophical meaning, ontology is the study of the foundations of what exists, what is common and most general among it, and what its origins are (see, e.g., Jacquette, 2002; Stróżewski, 2004, p.32). Hereafter, we refer to this concept by a boldface, capital "O" without a subscript (i.e., **O**).

Ontology in philosophy may also refer to "what exists" in a much more constrained, narrower, sense. This kind of ontology investigates existing, subject to a definition for existence, objects and relations in the world, and we will refer to this ontology as O. Depending on the assumptions made, different types of objects and relations may be recognized by O ontologies, because O branches into many

---

[5] Importing AI (or technical) aspects into philosophy brings with it a touch of reality that philosophical considerations often lack. See also the comment by Jacquette on the relation between a domain ontology and the domain itself (Jacquette, 2002, p.5).

[6] The term "being" is used in the sense employed by the Ancient Greeks, Parmenides (opposite to The Unbeing), Aristotle (Being qua being), medieval scholars like Aquinas (the study of being qua being) (Kerr, 2022), and some modern philosophers such as (Jacquette, 2002; Stróżewski, 2004; or Perzanowski, 2015). This term is also sometimes written as Being meaning totality of what exist (Kenny, 2012, p.160). Many modern philosophies, scientists computer engineers infuse this term with many different meanings (see this paper for the examples) obfuscating the original Greek sense of *onto-logia* – the fundamental study of being, probably as too esoteric (i.e., metaphysical) for their tastes (see also Kenny, 2012; Hofweber, 2021).

subdomains. Thus, many different perspectives have been developed for ontology (e.g., Quine, 1960; Jacquette, 2002; Stróżewski, 2004; Baker, 2007; Chalmers, Wasserman and Manley, 2009; Effingham, 2013; Ingarden, 2013; 2016; Berto and Plebani, 2015; Perzanowski, 2015; Thomasson, 2015; Hofweber, 2021); these differ in terms of extent, content, consistency, and accuracy, often responding to the specific needs of a domain.

In computational systems, ontology can be defined as "a specific vocabulary (dictionary) used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words" (see Guarino and Giaretta, 1995; Guarino, Oberle and Staab, 2009). In this context, ontologymay also refer to "a model of the structure of a system" (Guarino, Oberle and Staab, 2009) or "a formal, explicit specification of a shared conceptualization" (Studer, Benjamins and Fensel, 1998). We also have computational ontologies, which often called engineering ontologies, that "are machine-processable structures which represent particular domains of interest" (Husáková and Bureš, 2020). Ontology may also be used to refer to knowledge-based systems, databases, or AI systems that manage knowledge bases (see the discussions of Sharman, Kishore and Ramesh, 2007; Staab and Studer, 2009; Garbacz and Oliver Kutz, 2014; Husáková and Bureš, 2020).

The ontology of AI_E systems, meanwhile, denotes and determines how an AI_E system represents and reasons about the real world. Itis not concerned with computational representations of reality through structures, theories, data constructs, or computational concepts but rather with how real world objects, properties, and relations are registered by an AI system, as well as how they are recognized and interpreted. In brief, this ontology is solely committed to the real world (in the sense

explained by Smith, 2019, p.145), with it reflecting the real world[7], or physical reality, and the AI system's place in this world. More specifically, it is embodied, embedded, extended, and enactive (Smith, 2019, p.43), which are terms that will be explained later in this paper. There is no formal theory to accompany 4E ontology (ontology of AI system that is embodied, embedded, extended, and enactive), so there are no criteria for theoretical truth verification, but verification comes instead from a confrontation with the real world, which we will discuss later. A 4E ontology is not given in the form of a set of a priori relations and objects but rather acquired (Smith, 2019) in response to a dynamically changing reality (see Minsky, 1991, as well as; Bołtuć, 2020). We will refer to this ontology as O_E.

Let us now put these things together. Ontology (**O**), as a philosophical discipline, asks what is, in a most general sense, and what exists Ontology (O) is more restricted with the scope of this ontology being defined by the horizon of interest: For example, it may be the universe, some aspect of it, or a domain of reality (i.e., domain ontology), like ontology of biology, or ontology of physics. Ontology in computational systems, meanwhile, can be defined as "a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary entries," while the ontology of AI systems relates specifically to a representation of the

[7] The term "real world", or reality, is understood here as it is understood in Smith (2019, p.xiv), denoting the physical world we live in. Minsky (1991, p.6) refers to this reality as common sense reality. The term may be opposed to "virtual worlds", "imaginary worlds", "fantastic worlds", or other qualified uses of "words" denoting worlds as creations of computer systems, artistic expressions, or imaginations. The term "world" or "real world" may have multiple interpretations that we have no intention to discuss as such a discussion would be pointless and would not further the main point of the paper. Thus, the reader seeking more detailed explanation of this term should follow the cited references.

world or some knowledge domain in AI systems. Finally, the ontology of an AI_E system (O_E) refers to how AI_E system represents about the real world and how it is situated within reality.

Figure 1 illustrates these ontological dependencies by showing them in terms of their increased specificity both in scope as well as in application domain, from the most general (**O**), which is the ontology of what exists (O), to the most specific one, which in this case is O_E, the ontology of an AI_E system.

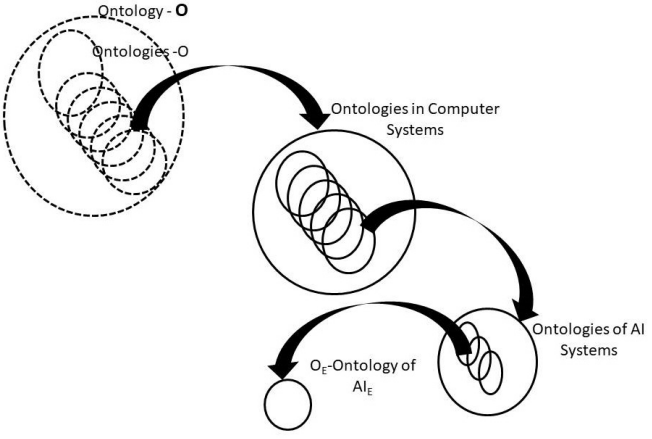

Figure 1: AI_E ontology and hierarchy of ontologies.

Thus, with respect to specificity and scope, O_E falls under the AI ontologies, which in turn are subspecies of the ontologies of computer systems, and these in turn are forms of specific ontologies (O) concerned with a specific segment of reality, which then falls under fundamental ontology (**O**) for existence. All these ontologies ask the same question but within different contexts, scopes, and perspectives. Furthermore, Figure 1 shows how only ontology **O** attempts to com-

prehend all that exists, with other ontologies being mere fragments. The further away we move from the fundamental ontology **O**, the narrower and more specific the scope of the ontology becomes.

Thus, ontology that represents the real world being always expressed in some form of language (which always is) is always incomplete with respect to the world, i.e., the world "as is" cannot be represented by something else, despite the close isomorphism between the reality and some logical aka ontological systems.[8] As a result, there will always be some inconsistency between what exists and what a given ontology represents, because ontology (even ontology **O**) will always be a theory (of sorts), expressed in some specialized language (see ft. 8) about the world rather than the world itself. Thus, there will always be a degree of incommensurability,[9] or a gap, between the ontology of synthetic systems and the reality of what exists. This gap may be narrowed but never entirely closed. Indeed, a representation can never contain itself as a part of what exists. We should keep this incommensurability in mind when building synthetic systems like AI_E.

The meta-ontology of AI

Meta-ontology is a relatively recently coined term with many interpretations. In philosophy meta-ontology denotes a study of ontology, what it investigates, and what it is concerned with (e.g., Quine, 1960;

[8] All pure ontological systems are logical systems (Jacquette, 2002, p.xiii; see e.g., Foschini, 2013).

[9] Incommensurability is not understood here in the same way as the incommensurability of paradigms or scientific theory in the works of Khun (1962) or Feyerabend (e.g., Ryan, 2002; for the incommensurability of paradigms, see also Sankey, 1993; Oberheim and Hoyningen-Huene, 2018; Bird, 2000). Instead, incommensurability is understood here as a general sense of not being entirely comparable according to some criteria.

Eklund, 2008; Berto and Plebani, 2015). Peter Van Inwagen (1998), the originator of the term, posited that the role of meta-ontology is to clarify the subject of ontology and explain how ontological claims can be interpreted. Francesco Berto and Matteo Plebani describe meta-ontology in terms of "'meta-X' as the inquiry on the central concepts and procedures of discipline X" (Berto and Plebani, 2015, p.13), where "X" refers to ontology in this case. Meta-ontology asks what a philosopher means when he asks ontological questions or questions about ontology (Eklund, 2008; Turner, 2014). Furthermore, meta-ontology is also concerned with ontological commitments and the truth conditions for a given ontological theory (Van Inwagen, 1998). From this perspective, meta-ontology would inquire as to what kind of "things" an ontological theory (i.e., ontology) is committed to.[10]

Ontological commitment and truth condition are key terms for defining meta-ontology, and they are critical for differentiating between ontology and meta-ontology (Turner, 2014). On the conceptual level, the ontological commitment denotes what kind of ontology (i.e., what exists, what is) a system or a theory is supposed to represent. Gibson (2009, p.631) states that "theory is ontologically committed to an object only if that object occurs in all the ontologies of that theory." Thus, ontology may be committed to the existence of numbers, planets, subatomic particles, ghosts, values, ethics, and so on, anything as long as these "things" are recognized in "all the ontologies of that theory" (e.g., Eklund, 2008). While ontology is about existence, "what is", the ontological commitment is about representation of "what is" (Smith, 1998) and the capacity to represent it.

Moreover, ontological commitment also includes verifying the criteria for this ontological commitment (i.e., verifying the truth of

[10] We do not discuss Quine's meta-ontology as being specific to Quine's concept of ontology that is not considered here.

its existential claims). Ontological commitment can be reduced to the simple claim that "A man is committed to the truth of whatever he asserts" (Searle, 1969, p.112).[11] For some, however, particularly professional ontologists, Searle's take on ontological commitment seems perfunctory, but in the AI context, it provides a simple (i.e., operational) means for judging the scope of an AI ontology. We do not dispute the robustness of Searle's claim on ontological commitment; here we take it as a guide in practical applications.

Truth conditions are what makes an ontology correct, such that a "theory is ontologically committed to an object only if that object occurs in all the ontologies of that theory" (Gibson, 2009, p.631). This statement was rephrased by Rayo (2007) into the following claim: "[...] for a sentence to carry commitment to Fs is for the sentence's truth to demand of the world that it contain Fs." Gibson's claim therefore comes from a philosophical perspective. It requires that a theory (of ontology) can be expressed in sentences (of any language), so this truth condition boils down to an agreement between the theory and the world in question (the correspondence theory of truth[12]), which may be the real world (as is the case of AI_E systems) or an imaginary constructs or virtual realities whatever the domain of ontology is.

---

[11] See the critique of Searle by Inwagen (1991).

[12] We are not going here into the discussion of the correspondence theory of truth or truth in general. We assume that for the systems (natural or artificial) "being in the world" (no Heideggerian connotations) there must be some relation, however tentative and limited holding between them and the real world they are immersed in, the relation of truth. This relation in a case of evolution and synthetic systems is verified by systems' operational success (see for example the discussion of truth relation in Bird, 2000). The correspondence theory of truth seems to trouble only philosophers of the anti-realist, skeptical persuasion, not computer scientists or evolutionary biologists (Bird, 2000). "Operational success" in applied sciences may be thought as playing the role of empirical proof in naturalized epistemology (Bird, 2000, p.263).

AI_E Paradigm

The term "AI paradigm" is used in many AI-related papers, discussions, and articles, with it carrying various meanings, so there is no agreement about what it should denote. In principle, authors simply define a "paradigm" in a way that suits their narrative, thus confirming Ian Hacking's prediction that the term would become banal following Kuhn's publication (Hacking, 2012). Thus, when we talk about AI paradigm we need to show precisely, to avoid misinterpretations, what we are talking about and how our definition of paradigm differs from, or is similar to, from other definitions.

Schopman (1986) suggests that AI has not developed a specific paradigm, claiming that "[. . .] no computational paradigm has yet been produced: there is no single generally accepted way to do AI" (Schopman, 1986, p.6). Čaplinskas (1998), however, defines three AI paradigms: the behaviorist paradigm, the agent paradigm, and the artificial life paradigm. Norvig (1992; 2002), meanwhile, associates the term with Lisp programming to express the paradigm of Artificial Intelligence Programming as being equivalent to a programming approach. Next, Cristianini (2014) distinguishes four AI paradigms: data-driven AI, statistical AI, knowledge-driven AI, and reasoning-and-search-based AI. Without much explanation as to why, Leary (2017) claims that the new Google AI paradigm is machine learning, while in a blog post titled "The AI Paradigm Shift," Richardson (2018) refers broadly to the AI paradigm as an approach to engineering AI systems, such as deep learning (DL), machine learning (ML), natural language processing (NLP), and robotics. (Hernández-Orallo et al., 2020, p.2522), meanwhile, claims that the concepts of AI paradigms

have been used to denote "broad families of technical or conceptual approaches: 'symbolic' vs 'connectionist', reasoning vs learning, expert systems vs agents."

Much like Richardson, Romero (2021) refers to deep learning and machine learning as AI paradigms, while Yalçın (2021) refers to symbolic (i.e., a human-readable, symbolic representation of problems, logic, and search) and sub-symbolic (i.e., an implicit representation derived from experience-based learning with no symbolic representation of rules and properties) representations as AI paradigms. Villar et al. (2021) hint at relating the AI paradigm to versions of machine-learning and deep-learning methods. Meanwhile, Luhach Kumar and Elçi Atilla (2021) in their book use the term "paradigm" in several places but in different contexts, with its meaning being variably associated with programming, the applications of AI to smart computational cyberspaces, or execution paradigms associated with computer hardware. Next, Joseph Makokha (2021) distinguishes two AI paradigms, namely an AI-based one for rule-following methods and another one based on artificial neural network constructs. The term "AI paradigm" is often used to simply denote a method for AI learning or knowledge acquisition, and this is how Yonguin Xu et al. (2021) use this term to denote deep-learning approaches, such as supervised, unsupervised, and reinforced learning.

Several studies have posited that the current AI systems use two broad, conceptual constructs, namely the symbolic and the sub-symbolic (e.g., Searle, 1998; Harvey, 2013; Neapolitan and Jiang, 2018; Mitchell, 2019; Smith, 2019; Cole, 2020; Russell and Norvig, 2020; Wooldridge, 2021). Thus, the symbolic paradigm reflects symbolic representations of a priori defined concepts that may be implemented in various programming environments, while the sub-symbolic paradigm relates to less clearly defined concepts, such as

the multi-dimensional probability weights on the connections within an artificial neural network. An approach under the sub-symbolic paradigm would therefore be implemented using one of the various machine learning (ML) technologies. These two paradigms can also be fused into a neuro-symbolic paradigm (e.g., Bader and Hitzler, 2005; Garcez, Gori et al., 2019; Garcez and Lamb, 2020; Kautz, 2022) that combines symbolic and sub-symbolic elements. As pointed out earlier, the term "AI paradigm" does not point to a single software or hardware solution (see Minsky, 1991; Russell and Norvig, 2020).

In this paper, we follow the example of Searle, Harvey, Neapolitan and Jiang, Smith, Cole, Wooldrige, and Russell and Norwig in using the term "AI paradigm" to denote a broad, conceptual construct that underlies AI systems. AI paradigm as is here defined does not imply a specific implementation. The AI paradigm therefore allows for multiple implementations, formal structures, representations, programming methods, and processing algorithms,[13] with these all belonging to a single paradigm.

The meta-ontology of AI_E systems

We have concluded that in philosophy, meta-ontology is the study of what ontology is all about. In other words, it is a study of, or about, ontology, as well as what it investigates; after Berto and Plebani "'meta-X' as the inquiry on the central concepts and procedures of discipline X", where "X" refers to ontology in this case (Berto

[13] We use the term algorithm in the general sense of a procedure that can be conceptualized and implemented in a computer. This use follows Knuth's definition of a computational method as being "A procedure that has all of the characteristics of an algorithm except that it possibly lacks finiteness may be called a computational method" (Knuth, 2005, p.5).

and Plebani, 2015, p.13). We also concluded that ontological commitment and truth conditions represent the key concepts of meta-ontology, and these terms are critical for differentiating between ontology and meta-ontology.

The meta-ontology of AI_E systems retains the core philosophical meaning (i.e., what ontology is committed to), but it diverges from the philosophical concept in the details, because it assumes the perspective of AI_E systems. When we say it "retains the core philosophical meaning," we mean that meta-ontology of AI_E systems is concerned with ontology, or is about ontology (i.e., the philosophical meaning of meta-ontology). Contrary to the use in philosophy, however, the meta-ontology of the AI_E focuses not on a theory of ontology, as meta-ontologies in philosophy do, but rather on what the AI_E represents (i.e., the real world).

We also said that we are concerned with studying the meta-ontology of the AI_E paradigm, which is the broad all-encompassing conceptual construct that underlies AI_E systems, rather than any particular realization of it. Indeed, we assume that most realizations of AI_E systems have the same foundational assumptions; what we denote as AI_E systems paradigm, so what we can conclude about the ontology of AI_E paradigm will also hold for its realizations. Different paradigms (from the one assumed here) of AI_E systems are logically possible. But we limit this discussion to assumptions, formulated by Smith, that AI_E systems should possess if they are to mimic humans' ability to cope with the real world (i.e., have human-level intelligence).

Summing this all up, the meta-ontology of the AI_E paradigm[14] is concerned primarily with what is (or about) the ontology of AI_E systems, what its ontological commitment is, and what its truth condition is.

To make our claims about the ontology of AI_E systems more specific, we employ Smith's postulates for the AI_E paradigm (Smith, 2019, p.44). As we mentioned earlier, Smith's AI_E paradigm is not purely ontological, but it does commit AI_E systems to certain ontology. Smith posited that for an AI system to match human-level intelligence, it needs to be embodied, embedded, extended, and enactive. More specifically, embodied means that an AI_E agent's representation of the real world accounts for its body's position, size, senses, and movement, such that the body plays a critical role in shaping the "mind" and its internal representation of the world (i.e., embodied cognition). Extended, meanwhile, implies that the AI_E agent's representation of the real world accounts for the AI_E system's mind and body as part of the cognition process (i.e., a co-creating ontology). (For more about the discussion of embodied and extended cognition, see, for example, the works of Varerla (1991), Clark and Chalmers (1998), Anderson (2003), Pfeifer and Iida (2004), Rupert (2009), Rowland (2010), Shapiro (2010), Wheeler (2011), Kiverstein (2018), Bermúdez (2020), and Paul (2021)). Next, the embedded condition refers to the AI_E system being aware of the context surrounding a situation, which should be accounted for in its ontology (e.g., Hutchins, 1995; Pouw, van Gog and Paas, 2014). Finally, being enactive means that an AI_E agent fully participates in actions, both in mind and body (e.g., Varela, Thompson and Rosch, 1991, p.175; Klein, Moon and Hoffman, 2006;

[14] The discussion of the meta-ontology of the AI_E paradigm is based on ideas proposed by Brian Cantwell Smith (1998; 2019) and the works of Minsky (1991), Dreyfus (2016), Mitchell (2019), Roitblat (2020), and Wooldridge (2021).

Froese and Ziemke, 2009; "the brain is conceived as participating in the action" Gallagher et al., 2013; Di Paolo and Thompson, 2017; Hutto and Myin, 2017; Newen, De Bruin and Gallagher, 2018; Smith, 2019; Newen, Bruin and Gallagher, 2020; Käufer and Chemero, 2021; Shapiro and Spaulding, 2021; "enacted AI" Shin, 2021; Hipólito and van Es, 2022).

Still, it is not obvious what ontology is implied by these requirements. As well as, it is not obvious, as to how we should translate the four features of this AI_E paradigm into meta-ontological requirements; Smith neglects to offer any suggestions here (Smith, 1998; Mitchell, 2019). Thus, we reformulated Smith's claims about the ontology of AI_E systems into four meta-ontological theses that appear to fill the ontological lacuna in his specifications. Indeed, they would seem to be necessary for AI_E systems to be embodied, embedded, extended, and enactive. They are:

T1. The ontological commitment of AI_E is to the real world, the world of a human agent.
T2. The truth condition of the ontology of AI_E is not consistency with ontological theory but rather the real world.
T3. The truth condition of AI_E is verified through the operational success of an AI_E system.
T4. The ontology of the AI_E paradigm must account for the dynamic environment of the real world.

(T1) The ontological commitment of AI_E is to the real world,[15] world of a human agent. The world for AI_E is the same reality that a human actor would exist in. We could say that AI_E ontology is a partial ontology as opposed to one that covers the entire world,

[15]The term explained earlier in the text. See footnote 7.

so it is about a state of affairs, by which we mean a local, temporal, dynamic (as described by Minsky) reality of the everyday world. This partial ontology does not attempt to create a comprehensive ontology of existence but rather account for what exists, together with the state of affairs,[16] in the part of the actual world that is relevant to an AI_E agent, we may say, agent-relevant ontology.

AI_E ontology is therefore not a theory about what exists, abstract objects, possible worlds, and maximal worlds (e.g., Forbes, 1992; Textor, 2021). Indeed, rejecting models or theories about the world may be beneficial, as Brooks suggests (in her ontology of everyday objects): "When we examine very simple level intelligence we find that explicit representations and models of the world simply get in the way. It turns out to be better to use the world as its own model" (Brooks, 1991). For instance, the ontology of the AI_E paradigmmay not have to account for subatomic particles, quantum physics, or imaginary objects, so it does not have to resolve Russell's table paradox (Russell, 1912); it does not have to account for these or similar objects as it is an ontology of everyday world we live in (it is Minsky's commonsense reality, or Baker's the world of ordinary things, or "the world of medium–sized objects" (Baker, 2007, p.18)).

Accordingly, the ontological commitment of the AI_E paradigm is whatever an AI_E system can assert about the world (i.e., the entities, relations between them, etc.) given its paradigm. AI_E systems face the real world, so they are committed to things within their context, and they need to recognize reality's features. Thus, the ontological

[16] The term "state of affairs" is used in the sense employed by Jacquette (2002).

commitment in AI_E systems that seek to mimic our own ontological commitment must be geared toward recognizing the reality with limited *a priori* suppositions.[17]

(T2) The truth condition of AI_E is not consistency with ontological theory but rather the real world. The truth condition of **AI_E systems** does not depend on theory, and it is not committed to the truth of a sentence because there are no sentences or collections of them, as there is no *a priori* ontological theory defining the ontology of **AI_E systems**. The truth condition of **AI_E systems**' ontology is therefore not consistency with a closed conceptual schema or ontological theory but rather with reality, with "what is the world like" (Smith, 2019, p.57).

The truth condition of the AI_E paradigm shares some similarities with the truth condition of philosophy, which states that "theory is ontologically committed to an object only if that object occurs in all the ontologies of that theory" (Gibson, 2009, p.631). Or, in and alternative formulation by Rayo (2007) rephrased as follows: "for a sentence to carry commitment to Fs is for the sentence's truth to demand of the world that it contain Fs." However, the truth condition of AI_E systems requires AI systems to "deal with reality as it actually is—not in the way our language represents it as being" (Smith, 2019, p.34), i.e., the truth condition of AI_E systems does not have to satisfy any sentences; understood as the ontological claims expressed in some form of language.

[17] See also Käufer and Chemero's (2021, p.220) discussion of Heideggerian AI, which is very similar to Baker's ontology of ordinary things (Baker, 2007) and a critique of the representational approach to the world.

(T3) The truth condition of $\mathbf{AI_E}$ systems is verified through operational success. The truth condition of **$\mathbf{AI_E}$ systems** is not concerned with computational representations of reality in the form of structures, data constructs, or computational concepts. Objects that are registered in **$\mathbf{AI_E}$ systems** follow constitutive regularities and norms, but these are not known beforehand but instead derived and learned from the real world as the basis of being (Smith, 2019, p.103). There is no theory to go with it, so there is no criterion for truth verification that references a theory, at least if we accept that this statement is not a theory in itself. As Baker says in an article about the metaphysics of ordinary things, "the ultimate test of a metaphysical theory is. . . pragmatic" (Baker, 2007, p.11).[18] In other words, the truth condition of AI_E systems is verified pragmatically, i.e., through operational success, because they are solely committed to the world and their actions within it (Smith, 2019, p.145). The precise meaning of operational success of engineering (including AI_E systems) or natural systems (living organisms) depend on the specific system the term "operational success" is applied to. In biological systems operational success (mostly) means survival and reproduction. In artificial systems operational success means fulfilling design objectives. It is not always obvious what is operational success even in engineering systems. For factory robots operational success is a well-defined task – like proper welding of a pin or similar. For autonomous vehicles operational success means (among other things) collision avoidance. For AI_E systems operational success is proper response/decision to situations. Of course operational success is much harder to evaluate in some cases than a welding of a pin; like it is in a case of the notorious trolley problem (see e.g., Cathcart, 2013). "Operational success" of

[18] In Baker's ontology, the pragmatic mode of verification has nothing to do with pragmatic theories of truth in philosophy.

the trolley problem is a subject of endless debates between engineers, philosophers and enthusiasts of AI probably with a limited chance of success as these groups talk past each other; philosophers see the trolley problem as ethical problems, engineers as engineering problem, and enthusiasts of AI are too emotionally engaged to be rational. As we mentioned earlier, registered/recognized objects in **AI_E ontology** create regularities and norms, but rather than being known a priori, they are derived and learned from the real world, and this provides the grounding for AI_E ontology.

(T4) The ontology of AI_E systems must account for the dynamic environment of the real world. The complex and dynamic nature of the AI_E domain was discussed by Minsky: "...the objects and activities of everyday life are too endlessly varied to be described by precise, logical definitions and deductions. Commonsense reality is too disorderly to represent in terms of universally valid axioms. To account for such variety and novelty, we need more flexible styles of thought, such as those we see in human commonsense reasoning, which is based more on analogies and approximations than on precise formal procedures" (Minsky, 1991, p.6).[19]

Thus, the reality that the ontology of the AI_E paradigm must represent is specific to a situation, because raw reality is too disorderly to represent through universally valid axioms. Indeed, the reality/ontology faced by AI_E is too complex and nuanced to be definable by a closed set of formal rules, and any attempt to do so would result

[19] Minsky is obviously not the first or only person to recognize the messiness of reality, but he is one of the few AI researchers that did so in the early years of AI technology (others include, for example, Dreyfus (2016), Wooldridge (2021), Smith (2019), Bołtuć (2020), Mitchell (2019), Roitblat (2020), and Käufer and Chemero (2021)).

in a combinatorial explosion (Inder, 1996, p.26). This combinatorial explosion barrier implies that the ontology of the AI_E paradigm must eschew any formal *a priori* decision-making procedures.[20]

We remain unsure about how to design AI systems that implement Smith's embodied, embedded, extended, and enactive ontology (e.g., Hoffmann and Pfeifer, 2018). Nevertheless, as biological agents do have embodied, embedded, extended, and enactive ontology suited to their specific living niche, we can assume that, in principle, synthetic systems could do the same, at least to some degree and perhaps with the use of technology that may not yet exist.[21]

Conclusions

In summation, the meta-ontological claims about AI_E systems posit that the ontological commitment of an $\mathbf{AI_E}$ system is directed solely to the outside world. In addition, the truth condition of $\mathbf{AI_E}$ **systems**' ontology is not consistency with a closed conceptual schema or ontological theory but rather with the reality, with "what is the world" (Smith, 2019, p.57). It is not concerned with computational representations of reality in the form of structures, data constructs, or computational concepts. In addition, this truth condition of $\mathbf{AI_E}$ **systems** is verified through operational success.

---

[20] Philosophical ontology recognizes that (to some extent) the needs of AI_E ontology seem to be the "ontology of everyday life" described by (Baker, 2007). Baker describes the ontology of common objects (i.e., the "metaphysics of everyday objects"), and this ontology may provide a philosophical interpretation for AI_E ontology, but possible similarities would again require further study.

[21] Smith's concept is similar to 4E cognition (e.g., Shapiro, 2010; Wheeler, 2011; Newen, Bruin and Gallagher, 2020). The field of 4E cognition requires a separate discussion because it lies outside the scope of this paper.

We are well aware of how natural systems engage successfully (most of the time) with the real world, and we know, at least in some sense, how they achieve this (see, for example, studies of 4E cognition (Shapiro, 2010; Wheeler, 2011; Newen, Bruin and Gallagher, 2020) or the work of Yong (2022)). To replicate the prowess of natural systems in synthetic systems, at least to some extent, we know that we need to mimic what natural systems do (i.e., engage with the real world (see e.g. Sarosiek, 2021)). In fact, we do not have any other example to follow but us and some other animals.

We also know that we must do something different to the way in which we approach AI systems now. In other words, we must change our AI paradigm, i.e., foundational assumptions about constructing AI systems (We refer again to the ideas of Minsky (1991), Dreyfus (2016), Wooldridge (2021), Smith (2019), Mitchell (2019), and Roitblat (2020).) Alas, we still do not know how to do this effectively.

We also know that we will always have a degree of incommensurability between the ontology of synthetic systems (including AI_E systems) and reality, which is in a sense explained in ft. 8, because there is also insurmountable incommensurability between the ontology of biological agents and reality. This means there will always be some shortfall between what exists and what can be comprehended by a system, whether biological or synthetic (e.g., Yong, 2022). Indeed, the ontology of a cognitive agent, whether natural or synthetic, always only partially covers reality (revisit Figure 1), because for biological systems, it is tailored to its environmental niche and continued survival, while for synthetic systems, it is oriented toward ensuring the utility of a system and the safety of those related to, on relaying on, this system. The best we can do is to minimize this incommensurability gap, once we realize that it exists, by optimizing a system to suit a specific environment.

Indeed, the meta-ontological lesson from nature is not that organisms strive to match their ontology with **O** ontology but rather to optimize their ontology to best meet their needs (see e.g., Yong, 2022) or occupy their biological niche, although this niche is essentially what their ontology is. Is this the way to go for AI_E systems? Obviously, a factory robot tightening nuts and bolts does not need an AI_E ontology, but a robot delivering pizza in a city would require a more sophisticated ontology. And so would police robots with the license to kill patrolling the city streets (see e.g., Propper, 2022). Furthermore, robotic companions, nurses, or personal assistants may require a still higher level of AI_E ontology. Such robots would need to navigate the messy environment of everyday life with the sort of cleverness that we expect from their human counterparts. In other words, they need human-level intelligence with human-level ontology.

The meta-ontological perspective, in the absence of generally accepted criteria, may also be useful for defining the AI paradigm, which could then be differentiated not by computing methods, software, or theories (as it is the case now) but rather by the ability to represent the real world. Such a perspective would clearly separate symbolic, sub-symbolic, or neuro-symbolic systems from their $\mathbf{AI_E}$ peers.[22]

[22] Meta-ontology has been used as a differentiating criterion between ontological paradigms. For example, Eklund (2008) uses meta-ontology to differentiate between ontological paradigms, such as between robust and deflationary conceptions of ontology.

Bibliography

Anderson, M.L., 2003. Embodied Cognition: A field guide. *Artificial Intelligence* [Online], 149(1), pp.91–130. https://doi.org/10.1016/S0004-3702(03)00054-7.

Bader, S. and Hitzler, P., 2005. *Dimensions of Neural-symbolic Integration – A Structured Survey* [Online]. arXiv. Available at: <http://arxiv.org/abs/cs/0511042> [visited on 4 January 2023].

Baker, L.R., 2007. *The Metaphysics of Everyday Life: An Essay in Practical Realism*. 1st ed., *Cambridge Studies in Philosophy*. Cambridge: Cambridge University Press.

Berto, F. and Plebani, M., 2015. *Ontology and Metaontology: A Contemporary Guide*. London; New York: Bloomsbury Academic.

Bird, A., 2000. *Thomas Kuhn*. 1st ed., *Philosophy now*. Chesham: Acumen.

Bołtuć, P., 2020. Conscious AI at the Edge of Chaos. *Journal of Artificial Intelligence and Consciousness* [Online], 07(01), pp.25–38. https://doi.org/10.1142/S2705078520500010.

Brooks, R.A., 1991. Intelligence without representation. *Artificial Intelligence* [Online], 47(1-3), pp.139–159. https://doi.org/10.1016/0004-3702(91)90053-M.

Čaplinskas, A., 1998. AI paradigms. *Journal of Intelligent Manufacturing* [Online], 9(6), pp.493–502. https://doi.org/10.1023/A:1008880017722.

Cassimatis, N.L., Bello, P. and Langley, P., 2008. Ability, Breadth, and Parsimony in Computational Models of Higher-Order Cognition. *Cognitive Science* [Online], 32(8), pp.1304–1322. https://doi.org/10.1080/03640210802455175.

Cathcart, T., 2013. *The Trolley, or, Would You Throw the Fat Man Off the Bridge?: A Philosophical Conundrum*. New York: Workman.

Chalmers, D.J., Wasserman, R. and Manley, D., eds., 2009. *Metametaphysics: New Essays on the Foundations of Ontology*. Oxford: Clarendon Press.

Clark, A. and Chalmers, D., 1998. The extended mind. *Analysis* [Online], 58(1), pp.7–19. Available at: <https://www.jstor.org/stable/3328150> [visited on 3 June 2019].

Cole, D., 2020. The Chinese Room Argument. In: E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy*. Winter 2020. Metaphysics Research Lab, Stanford University.

Di Paolo, E. and Thompson, E., 2017. *The Enactive Approach* [Online]. (preprint). MindRxiv. https://doi.org/10.31231/osf.io/3vraf.

Dreyfus, H.L., 2016. *Skillful Coping: Essays on the Phenomenology of Everyday Perception and Action*. Ed. by M.A. Wrathall. First published in paperback. Oxford; New York, NY: Oxford University Press.

Effingham, N., 2013. *Introduction to ontology*. Cambridge, UK; Malden, MA: Polity Press.

Eklund, M., 2008. The Picture of Reality as an Amorphous Lump. In: T. Sider, J. Hawthorne and D.W. Zimmerman, eds. *Contemporary Debates in Metaphysics*, *Contemporary debates in philosophy*, 10. Malden, MA: Blackwell Pub, pp.382–396.

Fjelland, R., 2020. Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications* [Online], 7(1), p.10. https://doi.org/10.1057/s41599-020-0494-4.

Forbes, G., 1992. Worlds and States of Affairs: How Similar Can They be? In: K. Mulligan, ed. *Language, Truth and Ontology* [Online]. Dordrecht: Springer Netherlands, pp.118–132. https://doi.org/10.1007/978-94-011-2602-1_8.

Foschini, L., 2013. *Where the "it from bit" come from?* [Online]. arXiv. Available at: <http://arxiv.org/abs/1306.0545> [visited on 4 January 2023].

Froese, T. and Ziemke, T., 2009. Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence* [Online], 173(3-4), pp.466–500. https://doi.org/10.1016/j.artint.2008.12.001.

Gallagher, S., Hutto, D.D., Slaby, J. and Cole, J., 2013. The brain as part of an enactive system. *Behavioral and Brain Sciences* [Online], 36(4), pp.421–422. https://doi.org/10.1017/S0140525X12002105.

Garbacz, P. and Oliver Kutz, 2014. *Formal Ontology in Information Systems*. Proceedings of the 8th international conference fois 2014, *Frontiers in artificial intelligence and applications*, v. 267. Washington, DC; Amsterdam: IOS Press.

Garcez, A.d., Gori, M. et al., 2019. *Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning* [Online]. arXiv. Available at: <http://arxiv.org/abs/1905.06088> [visited on 5 January 2023].

Garcez, A.d. and Lamb, L.C., 2020. *Neurosymbolic AI: The 3rd Wave* [Online]. arXiv. Available at: <http://arxiv.org/abs/2012.05876> [visited on 5 January 2023].

Gibson, R.F., 2009. Ontological Commitment. In: R. Audi, ed. *The Cambridge dictionary of philosophy*. 2. ed., 11. printing. Cambridge: Cambridge Univ. Press, p.631.

Goertzel, B., 2015. Artificial General Intelligence. *Scholarpedia* [Online], 10(11), p.31847. https://doi.org/10.4249/scholarpedia.31847.

Guarino, N. and Giaretta, P., 1995. Ontologies and Knowledge Bases: Towards a Terminological Clarification. In: N.J.I. Mars, ed. *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*. Amsterdam: IOS Press, pp.25–32.

Guarino, N., Oberle, D. and Staab, S., 2009. What Is an Ontology ? In: S. Staab and R. Studer, eds. *Handbook on Ontologies* [Online]. 2nd ed., *International Handbooks on Information Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp.1–20. https://doi.org/10.1007/978-3-540-92673-3.

Hacking, I., 2012. Introductory Essay. *The structure of scientific revolutions*. Fourth edition. Chicago; London: The University of Chicago Press, pp.6–30.

Harvey, I., 2013. Perspectives on Artificial Intelligence: Three Ways to Be Smart. In: I. Harvey et al., eds. *SmartData* [Online]. New York, NY: Springer New York, pp.27–38. https://doi.org/10.1007/978-1-4614-6409-9_3.

Haugeland, J., 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Hernández-Orallo, J. et al., 2020. AI Paradigms and AI Safety: Mapping Artefacts and Techniques to Safety Issues. *ECAI 2020* [Online]. Ed. by G. Giacomo, pp.2521–2528. https://doi.org/10.3233/FAIA200386.

Hipólito, I. and van Es, T., 2022. Enactive-Dynamic Social Cognition and Active Inference. *Frontiers in Psychology* [Online], 13, p.855074. https://doi.org/10.3389/fpsyg.2022.855074.

Hoffmann, M. and Pfeifer, R., 2018. Robots as powerful allies for the study of embodied cognition from the bottom up. In: A. Newen, L. De Bruin and S. Gallagher, eds. *The Oxford Handbook of 4E Cognition* [Online]. Oxford: Oxford University Press, pp.841–861. https://doi.org/10.1093/oxfordhb/9780198735410.013.45.

Hofweber, T., 2021. Logic and Ontology. In: E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy* [Online]. Spring 2021. Stanford, CA: Metaphysics Research Lab, Stanford University. Available at: <https://plato.stanford.edu/archives/spr2021/entries/logic-ontology/>.

Husáková, M. and Bureš, V., 2020. Formal Ontologies in Information Systems Development: A Systematic Review. *Information* [Online], 11(2), p.66. https://doi.org/10.3390/info11020066.

Hutchins, E., 1995. *Cognition in the Wild*. Cambridge, MA; London: MIT Press.

Hutto, D.D. and Myin, E., 2017. *Evolving Enactivism: Basic Minds Meet Content*. Cambridge, MA: MIT Press.

Inder, R., 1996. Planning and Problem Solving. In: M.A. Boden, ed. *Artificial Intelligence* [Online], *Handbook of Perception and Cognition*. San Diego: Academic Press, pp.23–53. https://doi.org/10.1016/B978-012161964-0/50004-2.

Ingarden, R., 2013. *Controversy Over the Existence of the World. Vol. 1* (A. Szylewicz, Trans.), *Polish Contemporary Philosophy and Philosophical Humanities*, 6. Frankfurt am Main [etc.]: Peter Lang Edition.

Ingarden, R., 2016. *Controversy Over the Existence of the World. Vol. 2* (A. Szylewicz, Trans.), *Polish Contemporary Philosophy and Philosophical Humanities*, 8. Frankfurt am Main [etc.]: Peter Lang Edition.

Inwagen, P.v., 1991. Searle on Ontological Commitment. In: E. LePore and R. Van Gulick, eds. *John Searle and His Critics*, *Philosophers and their critics*. Oxford: Blackwell, pp.345–358.

Jacquette, D., 2002. *Ontology* [Online]. Routledge. https://doi.org/10.4324/9781315710655.

Käufer, S. and Chemero, A., 2021. *Phenomenology: An Introduction*. Second edition. Cambridge, UK; Medford, MA: Polity Press.

Kautz, H.A., 2022. The third AI summer: AAAI Robert S. Engelmore Memorial Lecture. *AI Magazine* [Online], 43(1), pp.105–125. https://doi.org/10.1002/aaai.12036.

Kenny, A., 2012. *A New History of Western Philosophy*, *New History of Western Philosophy*. Oxford; New York: Oxford University Press.

Kerr, G., 2022. *Aquinas: Metaphysics*. Available at: <https://iep.utm.edu/thomas-aquinas-metaphysics/> [visited on 5 January 2023].

Kiverstein, J., 2018. Extended Cognition. In: A. Newen, L. De Bruin and S. Gallagher, eds. *The Oxford Handbook of 4E Cognition* [Online]. Oxford: Oxford University Press, pp.19–40. https://doi.org/10.1093/oxfordhb/9780198735410.013.45.

Klein, G., Moon, B. and Hoffman, R., 2006. Making sense of sensemaking 1: Alternative Perspectives. *IEEE Intelligent Systems*, 21(4), pp.70–73.

Knuth, D.E., 2005. *The Art of Computer Programming. Vol. 1: Fundamental Algorithms*. 3rd. Boston [etc.]: Addison-Wesley.

Krzanowski, R. and Polak, P., 2022. Ontology and AI Paradigms. *Proceedings* [Online], 81(1), p.119. https://doi.org/10.3390/proceedings2022081119.

Kuhn, T.S., 1962. *The structure of scientific revolutions. The structure of scientific revolutions.* University of Chicago Press: Chicago.

Leary, K., 2017. *Google's AI is a "new paradigm" that unites humans and machines*. Available at: <https://futurism.com/googles-ai-is-a-new-paradigm-that-unites-humans-and-machines> [visited on 5 January 2023].

Luhach, A.K., Elçi, A. and Sugumaran, V., eds., 2021. *Artificial Intelligence Paradigms for Smart Cyber-Physical Systems* [Online], *Advances in Systems Analysis, Software Engineering, and High Performance Computing*. IGI Global. https://doi.org/10.4018/978-1-7998-5101-1.

Lutkevich, B., 2022. *What is Artificial General Intelligence?* Available at: <https://www.techtarget.com/searchenterpriseai/definition/artificial-general-intelligence-AGI> [visited on 5 January 2023].

Makokha, J., 2021. Artificial Intelligence Paradigms and the Future of Learning: What a Partial Review of Half a Century of AI Conceptualiza-

tion Suggests. *2021 ASEE Virtual Annual Conference Content Access Proceedings* [Online]. Virtual Conference: ASEE Conferences. https://doi.org/10.18260/1-2--36700.

Marcus, G., 2022. *Artificial General Intelligence Is Not as Imminent as You Might Think*. Available at: <https://www.scientificamerican.com/article/artificial-general-intelligence-is-not-as-imminent-as-you-might-think1/> [visited on 5 January 2023].

Minsky, M.L., 1991. Logical Versus Analogical or Symbolic Versus Connectionist or Neat Versus Scruffy. *AI Magazine* [Online], 12(2), pp.34–34. https://doi.org/10.1609/aimag.v12i2.894.

Mitchell, M., 2019. *Artificial Intelligence: A Guide for Thinking Humans*. New York: Farrar, Straus and Giroux.

Neapolitan, R.E. and Jiang, X., 2018. *Artificial Intelligence: With an Introduction to Machine Learning, Second Edition* [Online]. 2nd ed. New York: Chapman and Hall/CRC. https://doi.org/10.1201/b22400.

Newen, A., Bruin, L.d. and Gallagher, S., eds., 2020. *The Oxford Handbook of 4e Cognition*. First published in paperback. Oxford: Oxford University Press.

Newen, A., De Bruin, L. and Gallagher, S., 2018. 4E Cognition: Historical Roots, Key Concepts, and Central Issues. In: A. Newen, L. De Bruin and S. Gallagher, eds. *The Oxford Handbook of 4E Cognition* [Online]. Oxford: Oxford University Press, pp.3–15. [visited on 5 January 2023].

Norvig, P., 1992. *Paradigms of Artificial Intelligence Programming: Case Studies in Common Lisp*. San Francisco, Calif: Morgan Kaufman Publishers.

Norvig, P., 2002. *A Retrospective on 'Paradigms of AI Programming'*. Available at: <https://norvig.com/Lisp-retro.html> [visited on 5 January 2023].

Oberheim, E. and Hoyningen-Huene, P., 2018. The Incommensurability of Scientific Theories. In: E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy* [Online]. Fall 2018. Stanford, CA: Metaphysics Research Lab, Stanford University. Available at: <https://plato.stanford.edu/archives/fall2018/entries/incommensurability/>.

Osaba, E. et al., 2021. *Artificial Intelligence - Latest Advances, New Paradigms and Novel Applications* [Online]. https://doi.org/10.5772/intechopen.87770.

Paul, A.M., 2021. *The Extended Mind: The Power of Thinking Outside the Brain*. Boston: Houghton Mifflin Harcourt.

Perzanowski, J.W., 2015. *Rozprawa ontologiczna i inne eseje*. Ed. by J. Sytnik-Czetwertyński. Toruń: Wydawnictwo Adam Marszałek.

Pfeifer, R. and Iida, F., 2004. Embodied Artificial Intelligence: Trends and Challenges. In: D. Hutchison et al., eds. *Embodied Artificial Intelligence* [Online]. Vol. 3139. Berlin, Heidelberg: Springer Berlin Heidelberg, pp.1–26. https://doi.org/10.1007/978-3-540-27833-7_1.

Pouw, W.T.J.L., van Gog, T. and Paas, F., 2014. An Embedded and Embodied Cognition Review of Instructional Manipulatives. *Educational Psychology Review* [Online], 26(1), pp.51–72. https://doi.org/10.1007/s10648-014-9255-5.

Propper, D., 2022. *San Francisco police proposal could allow cops to kill suspects with robots*. Available at: <https://nypost.com/2022/11/24/san-francisco-police-proposal-could-allow-cops-to-kill-suspects-with-robots/> [visited on 5 January 2023].

Quine, W.V., 1960. *Word and Object*. Cambridge, MA: The MIT Press.

Rayo, A., 2007. Ontological Commitment. *Philosophy Compass* [Online], 2(3), pp.428–444. https://doi.org/10.1111/j.1747-9991.2007.00080.x.

Richardson, F., 2018. *The AI Paradigm Shift*. Available at: <https://becominghuman.ai/the-ai-paradigm-shift-53fa07ae3ab2> [visited on 5 January 2023].

Roitblat, H.L., 2020. *Algorithms Are Not Enough: Creating General Artificial Intelligence*. Cambridge, MA: The MIT Press.

Romero, A., 2021. *Unpopular Opinion: We'll Abandon Machine Learning as Main AI Paradigm*. Available at: <https://towardsdatascience.com/unpopular-opinion-well-abandon-machine-learning-as-main-ai-paradigm-7d11e6773d46> [visited on 5 January 2023].

Rowlands, M., 2010. *The New Science of the Mind: From Extended Mind to Embodied Phenomenology* [Online]. The MIT Press. https://doi.org/10.7551/mitpress/9780262014557.001.0001.

Rupert, R.D., 2009. *Cognitive Systems and the Extended Mind* [Online]. 1st ed. Oxford; New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195379457.001.0001.

Russell, B., 1912. *The Problems of Philosophy* [Online]. New York : H. Holt. Available at: <http://archive.org/details/problemsofphilo00russuoft> [visited on 5 January 2023].

Russell, S.J. and Norvig, P., 2020. *Artificial Intelligence: A Modern Approach*. Fourth edition, global edition, *Pearson series in artificial intelligence*. Harlow: Pearson.

Ryan, J.G., 2002. *Feyerabend and incommensurability* [Online]. Master's thesis. Durham University. Available at: <http://etheses.dur.ac.uk/3754/> [visited on 5 January 2023].

Sánchez, D.M., Cavero, J.M. and Martínez, E.M., 2007. The Road Toward Ontologies. In: R. Sharman, R. Kishore and R. Ramesh, eds. *Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems* [Online]. New York: Springer, pp.3–20. Available at: <http://archive.org/details/springer_10.1007-978-0-387-37022-4> [visited on 5 January 2023].

Sankey, H., 1993. Kuhn's Changing Concept of Incommensurability. *The British Journal for the Philosophy of Science* [Online], 44(4), pp.759–774. https://doi.org/10.1093/bjps/44.4.759.

Sarosiek, A., 2021. The role of biosemiosis and semiotic scaffolding in the processes of developing intelligent behaviour. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* [Online], (70), pp.9–44. Available at: <https://zfn.edu.pl/index.php/zfn/article/view/535>.

Schopman, J., 1986. Artifical Intelligence and Its Paradigm. *Zeitschrift für allgemeine Wissenschaftstheorie / Journal for General Philosophy of Science* [Online], 17(2), pp.346–352. Available at: <https://www.jstor.org/stable/25170750> [visited on 5 January 2023].

Searle, J.R., 1998. *Mind, Language and Society: Philosophy in the Real World*. New York: Basic Books.

Searle, J.R., 1969. *Speech Acts: An Essay in the Philosophy of Language* [Online]. 1st ed. Cambridge University Press. https://doi.org/10.1017/CBO9781139173438.

Šekrst, K. and Skansi, S., 2022. Machine learning and essentialism. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (73), pp.171–196.

Shapiro, L., 2010. *Embodied Cognition* [Online]. London: Routledge. https://doi.org/10.4324/9780203850664.

Shapiro, L. and Spaulding, S., 2021. Embodied Cognition. In: E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy* [Online]. Winter 2021. Stanford, CA: Metaphysics Research Lab, Stanford University. Available at: <https://plato.stanford.edu/archives/win2021/entries/embodied-cognition/>.

Sharman, R., Kishore, R. and Ramesh, R., eds., 2007. *Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems* [Online]. New York: Springer. Available at: <http://archive.org/details/springer_10.1007-978-0-387-37022-4> [visited on 5 January 2023].

Shin, D., 2021. Embodying algorithms, enactive artificial intelligence and the extended cognition: You can see as much as you know about algorithm. *Journal of Information Science* [Online], p.016555152098549. https://doi.org/10.1177/0165551520985495.

Skuza, A., 2020. *What is artificial general intelligence and who builds it?* Available at: <https://arekskuza.com/the-innovation-blog/artificial-general-intelligence/> [visited on 4 January 2023].

Smith, B.C., 1998. *On the Origin of Objects*. 1st paperback, *A Bradford book*. Cambridge, MA: The MIT Press.

Smith, B.C., 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: The MIT Press.

Staab, S. and Studer, R., eds., 2009. *Handbook on Ontologies* [Online]. 2nd ed., *International Handbooks on Information Systems*. Berlin; Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-92673-3.

Stróżewski, W., 2004. *Ontologia*, *Kompendia Filozoficzne*. Kraków: Aureus; Znak.

Studer, R., Benjamins, V.R. and Fensel, D., 1998. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering* [Online], 25(1), pp.161–197. https://doi.org/10.1016/S0169-023X(97)00056-6.

Swar, A., Khoriba, G. and Belal, M., 2022. A unified ontology-based data integration approach for the internet of things. *International Journal of Electrical and Computer Engineering (IJECE)* [Online], 12(2), p.2097. https://doi.org/10.11591/ijece.v12i2.pp2097-2107.

Textor, M., 2021. States of Affairs. In: E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy* [Online]. Summer 2021. Stanford, CA: Metaphysics Research Lab, Stanford University. Available at: <https://plato.stanford.edu/archives/sum2021/entries/states-of-affairs/>.

Thomasson, A.L., 2015. *Ontology Made Easy*. Oxford: Oxford University Press.

Turner, J., 2014. Metaontology. In: Oxford Handbooks Editorial Board, ed. *The Oxford Handbook of Topics in Philosophy* [Online]. Oxford University Press, pp.1–18. https://doi.org/10.1093/oxfordhb/9780199935314.013.25.

Van Inwagen, P., 1998. Meta-ontology. *Erkenntnis* [Online], 48(2/3), pp.233–250. https://doi.org/10.1023/A:1005323618026.

Van Inwagen, P., 2009. *Metaphysics* [Online]. 3rd. Boulder, CO: Westview Press. Available at: <http://site.ebrary.com/id/10271914> [visited on 5 January 2023].

Varela, F.J., Thompson, E. and Rosch, E., 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge; London: The MIT Press.

Wheeler, M., 2011. Embodied cognition and the extended mind. In: J. Garvey, ed. *The Continuum Companion to Philosophy of Mind*, *Continuum companions*. London; New York: Continuum Press, pp.220–238.

Wooldridge, M., 2021. *The Road to Conscious Machines: The Story of AI*. London, UK: Penguin.

Xu, Y. et al., 2021. Artificial intelligence: A powerful paradigm for scientific research. *Innovation* [Online], 2(4), p.100179. https://doi.org/10.1016/j.xinn.2021.100179.

Yong, E., 2022. *An Immense World*. London: The Bodley Head.

Analysis of the implications of the Moral Machine project as an implementation of the concept of coherent extrapolated volition for building clustered trust in autonomous machines

Krzysztof Sołoducha
Military University of Technology, Poland

Abstract

In this paper, we focus on the analysis of Eliezer Yudkowsky's concept of "coherent extrapolated volition" (CEV) as a response to the need for a post-conventional, persuasive morality that meets the criteria of active trust in the sense of Anthony Giddens, which could be used in the case of autonomous machines. Based on the analysis of the results of the Moral Machine project, we formulate some guidelines for transformation of the idea of a coherent extrapolated volition into the concept of a coherent, extrapolated and clustered volition. The argumentation used in the paper is intended to show that the idea of CEV transformed into its clustered version can be used to build a technically and socially efficient decision-making pattern database for autonomous machines.

Keywords

ethics of artificial intelligence, ethics of autonomous machines, trust in artificial intelligence, moral machine project, coherent extrapolated volition.

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 231–255

The problem of the ethics of autonomous machines is a philosophical issue that has emerged together with the dangers of the development of modern technologies that allow the autonomisation of machine operation (Pasquale, 2016). Of course, the classic of considerations on this issue is Isaac Asimov (2004) and his reflections on the ethics of robots, but at the moment the discussion on this topic is determined primarily by the dynamic development of unsupervised machine learning methods (Gryz, 2021). They have gained a second wind (having been theoretically developed in the 1980s) thanks to the development of the Internet and access to large sets of data on which machines can learn and improve algorithms on their own, using formal rules of statistical reasoning and without any kind of supervision (Burrell, 2016).

This development brings certain risks. The most popular technologies based on the statistical paradigm, i.e. deep neural networks (more than 5 layers) or multilevel neural networks (such as Deepmind's AlphaGo), tend to have problems with a lack of transparency and explainability, which are currently the subject of intense discussion in the community related to the philosophy of artificial intelligence (Eschenbach, 2021). There are considerations concerning so called black box problem (Pasquale, 2016).

The classic problem of the ethics of robot activity raised by Asimov has thus been transformed today into a consideration of the concept of *benevolence*—the result of which is to build machines that do Good, especially under the threat of singularity. It means that machines are expected to act in favour of humans who are users of this technology and that machines shape their development in accordance with interest of mankind. It assumes the superiority of humans above machines and development of tools and methods which would be able to prevent bad scenarios like the threat of singularity—negative con-

sequences of faster development of machines than humans in terms of security. It is by the way very interesting topic that fundamental approach of humans towards machines is not to treat machines as equal to the humans (Karpus et al., 2021).

The solution to this issue, however, involves finding an answer to the fundamental question of whether machines that modify their algorithmic patterns based on statistical reasoning are able to recognise and reject those results of data analyses that lead to consequences that must be considered, despite their formal correctness, as ethically wrong. Autonomous machines should therefore have a mechanism to select and prevent the situations described, for instance, in the example of AI system malfunction presented below.

Asked for help by Ms Danni Morritt with a review of biology articles, Alexa—an artificial intelligence system developed by Amazon—suggested that if Danni would stab herself in the heart, it would reduce human pressure on the planet and save humanity from environmental catastrophe (Lo, 2019).

This, of course, is only one of many examples of reports on the problems of using AI systems, but it leads to a certain conclusion, which was, in fact, already utilised by Nick Bostrom in his well-known book entitled *Superintelligence* (Bostrom, 2016, p.306), and which we can slightly modify using Lawrence Kohlberg's theory of the development of ethical systems (Kohlberg, 1958; see also Górnicka, 1980; Czyżowska, Niemczyński and Kmieć, 1993). Kohlberg believes that human ethical development takes place in stages and that the highest level of development, the so-called level of universal principles, is reached by at most 20% of each population. In some ethical studies, for instance by Thomas Nagel (1986, p.208) such a level of development is also called *a third-person perspective*. Despite its limited representation in any population, this level of universal principles of

conscience is a reference point for other, less developed moral systems and determines the social effectiveness of behaviour that is the subject of ethical judgements.

However, the current stage of development of artificial intelligence based on statistical reasoning does not, in Bostrom's opinion, provide the possibility to reach such an ethical level in an autonomous way. It is very difficult from a technical point of view because the algorithms would have to be able to cross an individual utility calculation and apply abstract notions such as justice to optimise their performance. This difficulty can be clearly seen when we compare the ethical theories by Rawls and Nozick.

According to John Rawls' theory of justice (Rawls, 1971), the construction of such a third-person perspective requires the adoption of the minimax rule from the field of game theory, i.e. the abstract systemic perspective of developing a moral strategy for the worst possible situation in which a moral agent might accidentally situate himself. It is difficult to implement (Arrow, 1973; Harsanyi, 1975) and is criticised for example by libertarians as leading to harmful distributive outcomes that undermine systemic efficiency, and thus inconsistent with the notion of economic rationality developed within so-called classical economics (Wysocki, 2021)—the pursuit of maximising personal interests achieved in a systemically fair way (Nozick, 2013).

Using Bostrom's argument improved with the use of Kohlberg's theory, it is therefore possible to formulate the conclusion that the current statistical paradigm of artificial intelligence allows machines to reach, at most, the level of conventional morality created within the model of traditionally defined, classical rationality based on the maximisation of self-interest—it means the level of imitation of the behaviour of the majority of moral agents.

What is interesting—this argument is also supported by empirical research. According to a survey conducted in 2017 in the US, 78% of respondents declared a fear of using autonomous cars (Edmonds, 2017) and considered that their current development (implicitly statistical) does not inspire trust. This claim is supported by countless examples of bots that learn patterns of behaviour from available data downloaded from social media and create ethically incorrect patterns of automatic behaviour, based on statistical analysis of the data. The old argument about the difficulty of transition between the sphere of facts and the sphere of duty seems to be relevant here.

On the other hand, the arbitrary implantation of a certain abstract post-conventional ethic that breaks through these limitations raises the risk of being accused of usurpation and of acting with symbolic violence in terms of morality, which may also be met with lack of trust due to its arbitrary character. This lack of trust based on universalist violence can particularly occur in post-industrial and network societies, which are based on the so-called "active trust" model.

The concept of active trust

Active trust is a concept introduced by Anthony Giddens, the well-known English sociologist. According to Giddens, the problem of social trust involves providing a basic level of confidence to make rational decisions in a situation of uncertainty and lack of complete information. It is a permanent situation of a cognitive agent who does not have the status of an absolute, and it involves a trust-based reliance on individuals or abstract systems—based on a trust that balances ignorance or lack of information (Giddens, 1991, p.318). Giddens adds that in post-industrial and networked societies we are dealing with

so-called active trust, which is based on monitoring the honesty of the other person in an open and continuous way. Giddens' considerations on this subject can be supplemented in this regard by Fukuyama's approach, which perceives trust as an epiphenomenon of social capital and is a mechanism based on the assumption that other members of a community are characterised by honest and cooperative behaviour based on shared norms (Fukuyama, 1995, p.38). Sociological considerations about trust go much further, by proposing static and dynamic approaches and distinguishing different levels of trust (Miłaszewicz, 2016, pp.85–86). For the purposes of our deliberations, however, these subtleties do not seem noteworthy.

To sum up this stage of our considerations—from the point of view of the active trust that the operation of autonomy-based machines must generate, we accept Bostrom's argument that they cannot operate on the basis of a self-generated conventional morality, and that post-conventional morality in post-industrial societies cannot have the form of a universalistic usurpation.

So the problem of trustworthy autonomous machines can be reduced to the question of how to construct a model of post-conventional persuasive morality that meets Giddens' criteria.

Persuasive morality and the Moral Machine project

For the answer, we are going to use the distinction made by Virginia Dignum (2022). She identifies three possible approaches to the ethics of autonomous machines, distinguishing between ethics in technology design, which is to ensure that the ethical and social implications of these processes are taken into account in technology development processes; ethics by technology design, which is to ensure that, in the

case of autonomous machines, their automated reasoning processes contain correctly constructed ethical components; and ethics for technology designers ensuring the integrity of researchers and producers and the legal mechanisms that guide their work.

It is not difficult to guess that the concept of the automatic production of morality by machines, discussed above, covers the middle zone identified by Dignum. If, as we have shown above, this is impossible, then the question remains of how to create such a project of post-conventional morality of a persuasive nature, which would be instilled by system designers as a set of procedures and norms guiding the operation of autonomous systems as an external factor and not to be modified by machines working in a statistical paradigm.

Such an attempt was made by Eliezer Yudkowsky (2004) and he called this proposal a *coherent extrapolated volition* (CEV). This is a contemporary version of traditional virtue ethics, which is currently experiencing a renaissance due to its persuasive nature in the bottom-up model. This model is opposed to traditional top-down ethics, such as utilitarian ethics or deontological systems. However, the problem that is always related to the concrete implementation of virtue ethics is its local character, tied to the preferences and social practices of the particular community in which it is cultivated. Yudkowsky attempted to overcome this limitation by creating a programme of virtue ethics that would extend its reach not to the local community, but to the whole of mankind—meeting the universalist needs of the post-conventional model without relativistic limitation.

The idea of Coherent Extrapolated Volition is based on the concept of benevolent artificial intelligence, also proposed by Yudkowsky. It includes the following principles (Yudkowsky, 2004):

1. Benevolence—Artificial Intelligence (AI) must be friendly towards humans and all living beings and make choices that will be in the interest of everyone—third-person perspective.
2. Maintaining (preserving) benevolence—AI must want to pass on its value system to all its own descendants and instil these values in beings similar to itself.
3. Intelligence—AI must be smart enough to see how equality can be pursued through altruistic behaviour and try to do everything to make sure that the result of the undertaken action does not increase suffering.
4. Self-improvement—AI must feel the need and desire to continuously develop itself and to strive for such development among the surrounding living beings.

The notion *everyone*, which appears in the first principle of benevolence, was used by Yudkowsky to go one step further and propose a version of the third-person perspective that would not have local limitations. The proposal of the American researcher is declarative and based on the interpretation of the concept of extrapolation as statistical extrapolation. There is a certain paradox in this concept. Since we can take as its roots the negative assessment of the statistical foundations of contemporary autonomous systems as not offering any hope of producing post-conventional systems, it seems to be extravagant, to say the least, to use these tools to realise the project of contemporary virtue ethics. Yudkowsky's intention is the realisation of the eternal dream of constructing descriptive ethics that would deal with the problem of Hume's guillotine and show the path from facts to norms. Such a path would be statistical extrapolation, but realised on the scale of mankind.

The approach to extrapolation proposed by Yudkowsky turned out to be fruitful and can be taken as one of the inspirations for the creation of the Moral Machine project (Awad et al., 2020). In our study, we treat this project as a direct continuation of Yudkowsky's proposal. The second inspiration for this project, which appears directly in the references, is the concept of Indicators of Cultural Dimension (Hofstede, Hofstede and Minkov, 2010).

In this paper we make direct reference to the argument that inductive reasoning can be treated as a way of solving the is-ought problem. While this reasoning is fraught with the problem of uncertain inference, it is fundamentally consistent with Hume's inductive approach to, for example, the problem of causality. It is an approach that appeals to the weak rationality argument proposed by Searle (1964) and based on the concept of unreliability of purely logical reasoning about duty from description. From this point of view, one can also speak about an attempt to solve the so-called Jorgensen dilemma (Jörgensen, 1937) based on 3 claims:

- logically valid reasoning can be made only on the logical sentences (the ones, that can be true or false),
- the norms are not logical sentences,
- logical correct reasonings are carried out as practical syllogisms.

Therefore the facticity of the practical syllogism is based on reasoning grounded on weak rationality. And this concept of weak rationality used for moral reasoning was used in the Moral Machine project.

The Moral Machine project as an implementation of the idea of a coherent, extrapolated volition

The Moral Machine project was launched in 2014, being the result of collaboration between several academic centres (Exeter Business School, Massachusetts Institute of Technology, University of British Columbia, Max Planck Institute for Human Development, Toulouse School of Economics). Its aim was to gather via Internet as many opinions on moral dilemmas as possible, using as an example various modifications of the classic trolley model once proposed by Philippa Foot.[1] Using a special website (http://moralmachine.mit.edu), dilemma scenarios were presented to the public worldwide. The goal of the study was to identify solution patterns that could be used as a database for implementation in autonomous systems—the reference device in this case was an autonomous car.

39.61 million decisions from 133 countries were collected within the project and the decision databases were submitted to *conjoint analysis*. A conjoint analysis allows the study of the cumulative effect of specific characteristics of participants in a moral dilemma situation on moral preferences of cognitive agents making a decision in the face of a dilemma. The conjoint method is one of the methods of data classification and analysis that use a decomposition approach to measure the preferences of survey participants. Its core is to present a studied phenomenon as a particular combination of the features. These features are called attributes, and each attribute has a predefined number of levels. The identified attributes and their levels generate

[1] The issue of the value of the so-called trolley's dilemma for dealing with ethical problems is left here to be discussed in other contexts. Nevertheless, some arguments concerning this problem will be mentioned when discussing the critique of the Moral Machine project.

different variants, which are called profiles. The number of total profiles that can be generated depends on how many attributes and their levels we have (it is the multiplication of the number of levels of all attributes).

The Moral Machine study specifically searched for a quantity that is defined as the Average Marginal Component Effect (AMCE) of each of the moral situation attributes under study, i.e. the average effect of the characteristics of a particular attribute on the overall level of moral preference. In this way, there would emerge a Hofstede-like map of moral preferences.

Figures developed in the project show the nine AMCE values extracted from the data of the Moral Machine project. In each row of figures, the bar shows the difference between the probability of saving the character with the attribute on the right and the probability of saving the character with the attribute on the left, compared to the spread of all other attributes.

Nine attributes were identified that are taken as measures of preferences (and their opposites) of participants of the survey: intervention, relation to AV, gender, fitness, social status, law, age, no-characters, species. What is visible in the results of the analysis, the preferences to different degrees move in the direction of caring more about: inactivity rather than activity, concern for pedestrians rather than passengers, for females, for people in better physical shape and of a higher social status, following rules rather than breaking them, young versus old, using a utilitarian strategy in terms of calculating the amount of suffering, people versus animals. Moreover, for the different types of participants, it was discovered that, for example, people were preferred over animals and, among animals, dogs over cats. Among humans, on the other hand, children were preferred over adults.

The results of these research studies are interesting to the extent that they overlap to some degree and differ to some extent from, for example, the recommendations made *a priori* in 2017 by the *German Ethics Commission on Automated and Connected Driving*. For instance, there is complete overlap here in the preference for saving human lives at the cost of animals. On the other hand, the German recommendations are not clearly in favour of utilitarian strategies, while the mentioned above survey by Moral Machine project shows a clear preference for decisions based on quantitative criteria.

The greatest difference, however, occurs in the choices of certain features of participants in moral choice situations. German *a priori* rules would forbid gender or age preferences, and participants of the survey carried out in the Moral Machine project clearly show such.

There were also attempts to correlate the overall results with a precise, representative selection of 6 demographic indicators important for the entire survey population–age, education, gender, wealth, religion and political views. The analysis showed no significant differences in the results (the sample is then limited to 492,291 people).

Cultural clusters in the Moral Machine project

Interesting results have also emerged from an attempt to build cultural clusters in the manner of Hofstede's typology (Hofstede, Hofstede and Minkov, 2010). Geert Hofstede was a Dutch social psychologist and anthropologist who studied the effects of cultural differences on values. He developed a framework for understanding these cultural differences based on six dimensions: power distance, individualism,

masculinity, uncertainty avoidance, long term orientation, indulgence. Hofstede's cultural dimensions are a useful tool for understanding cultural differences (Hofstede, 2011).

With the help of geo-location technology, 130 countries with a representation of at least 100 respondents were selected. This resulted in a set of 448,125 survey participants. Using a clustering technique based on Euclidean metrics and Ward's method, three cultural clusters were identified—Western, Eastern and Southern. They generally coincide with the Ingelhart-Welzel map of cultural influences (Inglehart and Welzel, 2005).

The clusters are created as a result of the data analysis. They were integrated colourwise with Ingelhart-Welzel's map of cultural influences. There are significant differences between clusters in preferences for the 9 basic attributes of the survey. For instance, survey participants from collectivist cultures in the eastern cluster, where respect for elderly people is deeply rooted, showed less tendency to protect young people, as is typical, for instance, in the western cluster. Similar things occur, for example, regarding the attitude towards pedestrians who do not respect traffic regulations. In countries with a high organisational and legal culture from the western cluster, there is less tolerance towards such behaviour than in countries with less institutional traditions from the southern cluster. This also undermines, for example, the universality of German solutions in this area. In contrast, countries with high Gini index levels of social inequality tend to be more protective towards people with a higher social status, compared to those who are identified as coming from the lower reaches of society. Clustering, however, also made it possible to identify preferences that are very much cross-cultural. These are: protecting human life at the cost of animals, protecting many lives at the cost of fewer, and protecting young life.

The criterion of social mobility

The trend to look for cultural clustering in the process of choosing utilitarian strategies over deontological ones was inspiring further developed in a further publication by the authors of the Moral Machine project, entitled *Universals and variations in moral decisions made in 42 countries by 70,000 participants*. In this case, the differentiating indicator was the mobility index (Awad et al., 2020).

70,000 responses in 10 languages from 42 countries were selected for this project. A minimum of 200 responses from one country per scenario was assumed for the study. The split of the survey participants shows that there was a strong overrepresentation of European countries, the eastern coast of both American continents and some areas of Asia.

Many variants of the classical trolley dilemma were researched; these were called Switch, Loop, and Footbridge. The Switch scenario is a classic version of the trolley dilemma by Philippa Foot (2002). The moral agent has the ability to switch the path of the trolley so it would kill one person rather than five. This is a model situation for the application of a utilitarian strategy in which the mathematical summary of suffering counts and is the basis for decision-making in a dilemma situation.

In Loop scenario we deal with the active sacrifice of one life for the sake of five. The act of decision itself, however, does not result in direct killing. Indirect killing is faced if the man in blue on the bridge pushes the person next to him, that person will fall on the track. The trolley will hit that person and therefore not kill the five people working there. In Footbridge scenario we deal with the active sacrifice of one life for the sake of five linked with the act of direct killing.

Based on research by, among others, Joshua Greene (2013), it has been assumed that in the survey there would be expected a higher preference for the Switch and Loop over the Footbridge scenario, because research by moral psychologists shows that in the situation of necessity of direct killing there is a higher preference for death avoidance and the use of deontological strategies over utilitarian calculations. In the Switch and Loop scenarios, a less explicit distribution of preferences was assumed.

These assumptions were additionally correlated with the social mobility index, which was applied under the assumption that a high social mobility index allows for more behaviour that is socially unpopular and provides use of purely rational utilitarian strategies. In turn, low social mobility brings to the front limits and inhibitions that reduce the freedom to apply utilitarian models.

Another element that played a role in shaping the results of the survey were the cultural specifics of the different countries. For Asian countries, lower social mobility is also correlated with a lower propensity to express controversial opinions and to come into disagreement with the environment. This is indicated also, for example, by Hofstede's research.

According to the survey results, in the case of European countries and those from both American continents, there is a clear preference for the choice of utilitarian strategies. We can also observe much less inhibition to seek solutions based on utilitarian criteria. In the case of Asian countries, due to their cultural characteristics, there is generally a stronger tendency to be inhibited towards utilitarian ethics and a much stronger tendency to behave according to fears of the opinion of the surrounding community blaming the moral agent for behaviour incompatible with the social deontological taboo prohibiting intentional killing.

The cognitive value of the trolley model

The results of the *Moral machine* project presented above provoked much criticism. It is expressed, for example, in a critical article devoted to the inequalities uncovered in the study entitled *Life and death decisions of autonomous vehicles*, which was published in "Nature" (Bigman and Gray, 2020). The main criticism of the authors concerns the methodology used by those responsible for the *Moral machine* programme, which, in their view, is completely inadequate to deal with the problem of inequality. It concerns in particular the use of the ethical dilemma schema to study moral preferences. It forces a situation to be resolved unambiguously by choosing one of the ethical strategies, which ends up sacrificing one option to another (one death for another death). With such a construction of the dilemma, unequal treatment of the actors of the dilemma is forced—for instance women in favour of men. But when the equality option is added, e.g. between men and women, it is selected in more than 97 cases as the preferred option (a study of a competing version of the dilemma was performed on a group of approximately 1,000 Americans and 1,000 British people). According to the authors of the polemical statement, the preference for unequal treatment discovered during the MM project—taking into account the decisions by race, gender, age of the moral agents—should therefore not be taken into consideration when constructing action patterns for autonomous machines.

Such empirical results should be ignored in favour of a normative stance that prefers an egalitarian approach and the survey questions should be structured according to this assumption.

The authors of the MM project in their response, posted parallel to the critique, pointed out that in many of the survey elements it is possible to find non-preference options in favour of one of the

solutions, which means an egalitarian attitude. This is the case, for example, with gender preferences, the fitness of the actors or tendency to protect passengers or pedestrians.

Another reaction to the Moral Machine study is also a criticism of the whole model of using moral dilemmas to study moral preferences. This is due to the very nature of the dilemma, which is a specially constructed situation that has no good solution and requires the choice of some moral strategy to justify the choice of the "lesser evil". According to the arguments contained, for example, in the text *Trolled by trolley* (Mirnig and Meschtscherjakov, 2019) or in other studies (Holstein and Dodig-Crnkovic, 2018), research should focus on designing machines in such a way that they can rather anticipate and avoid dilemma situations than deciding who to kill at any given time in a dilemma situation.

The reference cluster problem

The idea of contemporary virtue ethics as clustering databases that underline decision-making of machines is connected to the problem of choosing a reference cluster for the operation of an autonomous machine at a particular place and time.

This problem essentially can be reduced to the question of whether, in the case of regionalisation associated with the clustering of virtue ethics, the machine should take into account the decision-making preferences of the driver and his own cultural cluster or the environment in which he travels—so the machine should navigate according to the rules of the territory in which it operates as a transportation tool.

In the case of a legal judgment, the situation is quite clear. The foreigner is bound by the law of the country of destination. But what about the problem of trust?

In order to find answers to this question, we carried out a sample survey on 34 students aged between 18 and 39, who were asked about their preferences in this respect. From a methodological point of view, this is a survey carried out in the form of an online questionnaire. Its aim was to examine the basic preferences of possible users of autonomous machines in terms of the expected level of clustering. The survey has no ambition to be a representative poll. Our objective is rather to identify certain trends in user preferences.

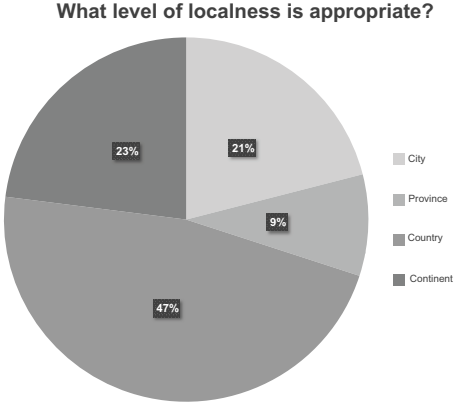

Figure 1: Most interesting results of the survey in terms of clustering of preferences of AV users. Source: Author's study.

The responses are split almost equally between the preferences of the driver and the preferences of the residents of the area in which the vehicle travels, with a slight advantage to the driver. The conclusion for manufacturers of autonomous devices is therefore that a device should have an open architecture that allows it to be adapted to the

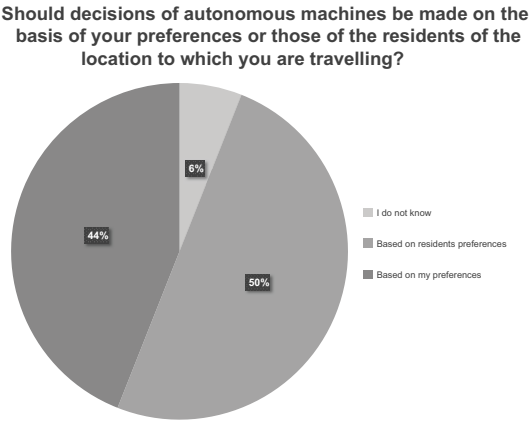

Figure 2: Level of clustering survey. Source: Author's study.

preferred option, unless legal regulations decide otherwise. However, as the survey results show, imposing solutions in this area may end in reduced trust in the device.

The other problem investigated in the survey was the level of clustering. The top level used in the research was the continental one. The majority of respondents preferred the level of state. The level of clustering reflecting Ingelhart-Welzel's map of cultural influences used in the research mentioned above did not appear among the questions. This does not reflect traditional geographical and administrative distinctions.

Conclusion

Despite the sceptical voices raised against the practical implementation of the idea of an extrapolated, coherent volition of humanity in the form of the Moral Machine project, it does not seem that this

criticism undermines certain important arguments that stand behind Yudkowsky's CEV project and its implementation attempts. In conclusion, we would like to point them out and outline some possible paths for further consideration of this issue.

Firstly, Yudkowsky's project and attempts to implement it are answers to the question of what moral patterns should be introduced into the decision-making procedures of machines so that they meet Giddens' active trust requirements. Any attempts to do this in a top-down model by enlightened bodies, special committees or top-down adopted normative systems does not meet the criterion of persuasive trust proposed by Giddens. Therefore, the attempt to do this in a descriptive way through empirical research referring to extrapolation of results of the survey seems a method tailored to these needs, with all of the doubts associated with the naturalisation of morality and the limitations of descriptive research related to Hume's guillotine problem. Although, in our paper, we tried to show that the weak rationality associated with the inductive approach applied to moral problems can be used to overcome the is-ought problem in the case of the ethics of autonomous machines.

Secondly, such considerations can be conducted under the assumption that the best path to their implementation is to define cognitive processes (including those of a moral nature) as consisting of information processing. What may help here may be the concept that such a definition of moral cognition would not be a naturalisation (Peruzzi, Aseron and Bhaskaran, 2015). This also could help to eliminate the troubling issue of the *is-ought problem*. Although it is an issue considered by some theorists to be an illusory (Gellner, 2005) and an argument that is treated as untenable nowadays (Searle, 1964) from point of view of "weak" rationality—as mentioned above.

Thirdly, the acceptance of the possibility of clustering moral patterns can lead to the idea of constructing the architecture of autonomous machines as open for adaptation of the patterns used for decision-making to the local characteristics of the user/social environment. The sample survey presented at the end of the paper demonstrates the preferences of a limited group of respondents regarding this issue.

Fourthly, there is quite advanced research on the technology of so-called social robots, whose task is to produce a personalised interactive communication experience by considering the preferences of the user the robot interacts with (Maroto-Gómez et al., 2022). It is based on technology so called preference learning (Fürnkranz and Hüllermeier, 2011). Using an online survey, participants provide their defining features and preferences towards the activities of the robot. Then, a preference learning model estimates the preferences of new users using similar features of the survey participants. The survey contains questions about sociodemographic, habits, interests, and preferences about specific attributes related to social robot (Fürnkranz and Hüllermeier, 2011, p.2).

Acknowledgments. The author wishes to thank anonymous helpful referees whose insightful comments helped improve the quality of the present paper. Most certainly, if there are still some errors remaining, they are my own responsibility.

Bibliography

Arrow, K.J., 1973. Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice. *The Journal of Philosophy* [Online], 70(9), pp.245–263. https://doi.org/10.2307/2025006.

Asimov, I., 2004. Runaround. *I, robot*. Bantam hardcover edition, *Bantam spectra book*. New York: Bantam Books, pp.30–55.

Awad, E. et al., 2020. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences* [Online], 117(5), pp.2332–2337. https://doi.org/10.1073/pnas.1911517117.

Bigman, Y.E. and Gray, K., 2020. Life and death decisions of autonomous vehicles. *Nature* [Online], 579(7797), E1–E2. https://doi.org/10.1038/s41586-020-1987-4.

Bostrom, N., 2016. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Burrell, J., 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* [Online], 3(1), pp.1–12. https://doi.org/10.1177/2053951715622512.

Czyżowska, D., Niemczyński, A. and Kmieć, E., 1993. Formy rozumowania moralnego Polaków w świetle danych z badania metodą Lawrence'a Kohlberga. *Kwartalnik Polskiej Psychologii Rozwojowej*, 2(1), pp.19–38.

Dignum, V., 2022. *Responsible AI: From Principles to Action*. Available at: <https://www.youtube.com/watch?v=LwKDOwwJpL4> [visited on 11 January 2023].

Edmonds, E., 2017. *Americans Feel Unsafe Sharing the Road with Fully Self-Driving Cars*. Available at: <https://newsroom.aaa.com/2017/03/americans-feel-unsafe-sharing-road-fully-self-driving-cars/> [visited on 11 January 2023].

Eschenbach, W.J.v., 2021. Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology* [Online], 34(4), pp.1607–1622. https://doi.org/10.1007/s13347-021-00477-0.

Foot, P., 2002. The Problem of Abortion and the Doctrine of the Double Effect. In: P. Foot, ed. *Virtues and Vices: And Other Essays in Moral Philosophy* [Online]. Oxford: Oxford University Press, pp.5–15. https://doi.org/10.1093/0199252866.003.0002.

Fukuyama, F., 1995. *Trust: The Social Virtues and the Creation of Prosperity*. 1st ed., *A Free Press paperbacks book*. New York: Free Press.

Fürnkranz, J. and Hüllermeier, E., 2011. Preference Learning: An Introduction. In: J. Fürnkranz and E. Hüllermeier, eds. *Preference Learning* [Online]. Berlin; Heidelberg: Springer, pp.1–17. https://doi.org/10.1007/978-3-642-14125-6_1.

Gellner, E., 2005. *Words and Things: An Examination of, and an Attack on, Linguistic Philosophy*. 1. publ, *Routledge classics*. London: Routledge.

Giddens, A., 1991. *Modernity and Self-Identity: Self and Society in the Late Modern Age*. Stanford, CA: Stanford University Press.

Górnicka, J., 1980. Rozwój moralny w koncepcji Lawrence'a Kohlberga. *Człowiek i Światopogląd*, 6, pp.113–123.

Greene, J.D., 2013. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. London: Atlantic Books.

Gryz, J., 2021. *Sztuczna Inteligencja: powstanie, rozwój, rokowania*. Available at: <https://www.youtube.com/watch?v=3ZDfVgC897k> [visited on 11 January 2023].

Harsanyi, J.C., 1975. Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory. *The American Political Science Review* [Online], 69(2). Ed. by J. Rawls, pp.594–606. https://doi.org/10.2307/1959090.

Hofstede, G.H., Hofstede, G.J. and Minkov, M., 2010. *Cultures and Organizations: Software of the Mind: Intercultural Cooperation and Its Importance for Survival*. 3rd ed. New York: McGraw-Hill.

Hofstede, G.J., 2011. *Geert Hofstede on Culture*. Available at: <https://www.youtube.com/watch?v=wdh40kgyYOY> [visited on 11 January 2023].

Holstein, T. and Dodig-Crnkovic, G., 2018. Avoiding the intrinsic unfairness of the trolley problem. *Proceedings of the International Workshop on Software Fairness* [Online]. Gothenburg Sweden: ACM, pp.32–37. https://doi.org/10.1145/3194770.3194772.

Inglehart, R. and Welzel, C., 2005. *Modernization, Cultural Change, and Democracy: The Human Development Sequence* [Online]. Cambridge: Cambridge University Press. https : / / doi . org / 10 . 1017 / CBO9780511790881.

Jörgensen, J., 1937. Imperatives and logic. *Erkenntnis* [Online], 7(1), pp.288–296. https://doi.org/10.1007/BF00666538.

Karpus, J. et al., 2021. Algorithm exploitation: Humans are keen to exploit benevolent AI. *iScience* [Online], 24(6), p.102679. https://doi.org/10.1016/j.isci.2021.102679.

Kohlberg, L., 1958. *The Development of Modes of Moral Thinking and Choice in the Years 10 to 16* [Online]. PhD thesis. Chicago, IL: University of Chicago. Available at: <https://www.proquest.com/openview/c503bf59d762abe5818e1b24c484d41a/1?pq-origsite=gscholar&cbl=18750&diss=y> [visited on 11 January 2023].

Lo, T., 2019. *"My Amazon Alexa Went Rogue and Ordered Me to Stab Myself in the Heart"*. Available at: <https://www.mirror.co.uk/news/uk-news/my-amazon-echo-went-rogue-21127994> [visited on 11 January 2023].

Maroto-Gómez, M. et al., 2022. An adaptive decision-making system supported on user preference predictions for human–robot interactive communication. *User Modeling and User-Adapted Interaction* [Online], pp.1–45. https://doi.org/10.1007/s11257-022-09321-2.

Miłaszewicz, D., 2016. Zaufanie jako wartość społeczna. *Studia Ekonomiczne* [Online], (259), pp.80–88. Available at: <http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.cejsh-d64b921a-5db7-4208-9eb7-73baaa05f7e4> [visited on 11 January 2023].

Mirnig, A.G. and Meschtscherjakov, A., 2019. Trolled by the Trolley Problem: On What Matters for Ethical Decision Making in Automated Vehicles. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* [Online]. Glasgow: ACM, pp.1–10. https://doi.org/10.1145/3290605.3300739.

Nagel, T., 1986. *The View from Nowhere*. 1st ed. Oxford: Oxford University Press.

Nozick, R., 2013. *Anarchy, State, and Utopia*. New York: Basic Books.

Pasquale, F., 2016. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

Peruzzi, N., Aseron, R. and Bhaskaran, V., 2015. *A Beginner's Guide to Conjoint Analysis*. Available at: <https://www.youtube.com/watch?v=RvmZG4cFU0k> [visited on 11 January 2023].

Rawls, J., 1971. *A Theory of Justice* [Online]. Cambridge, MA: The Belknap Press of Harvard University Press. Available at: <http://www.gbv.de/dms/bowker/toc/9780674880146.pdf> [visited on 11 January 2023].

Searle, J.R., 1964. How to Derive "Ought" from "Is". *The Philosophical Review* [Online], 73(1), pp.43–58. https://doi.org/10.2307/2183201.

Wysocki, I., 2021. The problem of indifference and homogeneity in Austrian economics: Nozick's challenge revisited. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* [Online], (71), pp.9–44. Available at: <https://zfn.edu.pl/index.php/zfn/article/view/554> [visited on 24 January 2022].

Yudkowsky, E., 2004. *Coherent Extrapolated Volition*. San Francisco, CA: Singularity Institute for Artificial Intelligence.

Essays

Eseje

Will a human always outsmart a computer? An essay

Adam Olszewski
Pontifical University of John Paul II in Kraków, Poland

Abstract

The title question of the paper has its empirical origin in the form of an individual's existential experience arising from the personal use of a computer, which we attempt to describe in the first section. The rest of the entire paper can be understood as a philosophical essay answering the question posed. First the connection between the main problem of the article and its "premonition" by mankind, which was expressed in the form of ancient myths and legends, is briefly suggested. After shortly discussing the problems that early considerations of AI focused on, i.e. whether machines can think at all, we move on to reformulate our title question, about the possibility of outsmarting AI. This outsmarting will be understood by us in a rather limited way as to prevent a machine from completing its implemented task. To achieve this objective, after softly clarifying the basic terms, an analogy is built between the "outsmarting" of a machine by a human (the target domain) and the playing of a mathematical game between two players (the base domain), where this outsmarting is assigned a "winning strategy" in the certain game. This mathematical model is formed by games similar to Banach-Mazur games. The strict theorems of such games are then proved and applied to the target of the analogy. We then draw conclusions and look for counter-examples to our findings. The answer to the title question posed is negative, and it is not clear how far it should be taken seriously.

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 259–280

Keywords
existential experience, myth, computer, machine, Banach-Mazur games, winning strategy.

> *Much, if not all of the argument for existential risks from superintelligence seems to rest on mere logical possibility* (Dubhashi and Lappin, 2017).

0. An existential ambient of the problem

The appearance of personal computers—which took place in the middle of 1970s in the world (USA) and a decade later in Poland (due to the "iron curtain" which at that time separated this country from the rest of the world)—brought hope for providing man with a useful tool with versatile applications and capable of performing certain very practical functions. Thus, in those days, from the perspective of an average user, a computer standing on his desk was a *machine*, that is a concrete, experiential and *human-friendly* device, supporting his activities, for example as a typewriter or a memory bank for storing data. Since then, over nearly fifty years, such a notion of a useful machine has considerably evolved. This has happened thanks to the extremely dynamic development of information technology (which included increasing the computational capabilities of machines, their speed and memory and developing new and better programming techniques and languages) and the emergence of the Internet. It seems that the present concept of a *machine* has undergone a substantial modification. This machine, which was a friendly tool supporting man and operating in virtual (non-real) time and sepa-

rated from the outside world, suddenly became a machine operating in real time and additionally started drawing us deeper and deeper into a kind of addiction and making us undertake certain activities, which—although performed through a machine standing on the desk—have far-reaching consequences in real time and, more importantly, in reality. This new machine on our desk is, in fact, a terminal of an extremely vast network (*inter-net*), in which powerful algorithms create the elements of the *machine proper*. This *new machine*, which has taken on *a new meaning*, is also interactive in a double sense. First, it is able to learn in a certain sense strictly defined by computability theory, and second, it is known that it must be managed by or is in the hands of a person or a group of people because it is not an independent entity (or at least we do not know about it yet). We can call such a machine *artificial intelligence* (AI). We deliberately omit here a wide array of issues that need to be clarified or systematically listed[1]; these shortcomings justify using the term *essay* in the title of this work. Besides, this paper is my first attempt at analyzing this somewhat strange subject, so I consider it to be the first of a series of papers and an essayistic outline of my further work on the issue.

1. The difference between an algorithm, a program, a machine, and AI

For the sake of clarifying the approach adopted, it will be good to describe succinctly and in a popular scientific way how we will understand the differences between an algorithm, a program, a machine and AI. The main philosophical issues in the relationship between an algorithm, a program, and a machine are:

[1] For some clarifications see (Krzanowski and Polak, 2022).

- the existence and mode of existence of an algorithm;
- the existence and mode of existence of programs;
- mutual relations between an algorithm and a program;
- the implementation of an algorithm of a program on a machine.

For the sake of further considerations, We will adopt a solution which refers to the solution of the problem of universals formulated by St. Thomas Aquinas. One *universal* (or idea) may exist and take one of three forms, depending on the way it is approached:

- *ante rem*—in the world of Platonic ideas—an algorithm;
- *post-ante rem*—as a construction of the mind—a program;[2]
- *in re*—physically present in the thing—implemented into a machine (computer) in the form of the processes that control the behavior of the physical device.[3]

In other words, the three objects under consideration are the same object which manifests itself in three different forms or phases. We are not necessarily going to defend the above distinction at all costs, but we believe that it is easy to understand and will facilitate further reasoning although it will not be essential for the core of our considerations.

Since artificial intelligence (AI) is an artefact that exists in reality, its definition should have the character of a *real definition*. However,

[2] *Post-ante* is a strange term which is to indicate that from one point of view a program appears before the *in re* phase and after the *ante rem* phase, while from another point of view it appears after the *in re* phase.

[3] Implementation is an extremely interesting concept in the philosophy of computer science. It allows us to transform something abstract into something physical, which requires thorough consideration. There is a certain similarity here with the criticism of Platonism, where one of the objections against Platonism is the impossibility to move from the abstract to the physical.

due to the fact that AI was brought into existence not so long ago, and the fact that its properties are still subject to almost constant fluctuations, it is difficult to give it an adequate real definition. As a result, most often the definitions given by authors of AI are the result of his/her individual decision and/or consist in choosing from a range of detailed and rigorious definitions available in the literature, which we will also do below. On the other hand, the property of intelligence that we attribute to an artificial system is of natural origin. After all, in the original sense, intelligence, as a property, is attributed to man or, by analogy, to an animal. Defining this property in humans is also of a reporting nature. According to researchers of the issue, from a methodological point of view, the general notion of goal is crucial for the definition of AI[4], as the following table summarises (Bringsjord and Govindarajulu, 2022; cf. Russell and Norvig, 2021):

	Human-Based	Ideal Rationality
Reasoning-Based:	Systems that think like humans.	Systems that think rationally.
Behavior-Based:	Systems that act like humans.	Systems that act rationally.

Table 1: Four Possible Goals for AI According to (Russell and Norvig, 2021).

For our purposes, the following definitions of AI are of particular value by pointing to this artificial component of AI as a *machine* or *computer system* or simply a *computer*; and presupposing the relevant *programmes* (algorithms) implemented in them.

[4] More about this see e.g. (Dodig-Crnkovic, 2022).

> The capacity of computers or other machines to exhibit or simulate intelligent behaviour; the field of study concerned with this.[5]
>
> Artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems. Specific applications of AI include expert systems, natural language processing, speech recognition and machine vision.[6]

The above definitions of AI are schematically based on the following *quasi-equalities*:

- E_1: AI = (System computer + Simulation + Data).

However, some progress has been made in recent years and has taken the form of so-called *adaptive systems*. For them, the quasi-equality looks slightly different:

- E_2: AI = (System computer + Simulation + Data + Interaction with Data).[7]

These two variants of AI are not, it seems, functionally equivalent, but it is clear from their presentations that E_1 is 'part of' E_2.[8] Given this important distinction, the reader should be forewarned that our considerations will concern E_1, and consequently AI will be understood abstractly as a single algorithm. However, to reassure the reader, let us note that:

---

[5] Cf. *Oxford English Dictionary*, entry: „artificial intelligence" (*Artificial Intelligence*, 2022).

[6] Cf. (Burns, Laskowski and Tucci, 2022). One can also add other goals here like: decision-making, translation between languages, visual perception; and others.

[7] „Adaptive AI can change its own code to incorporate what it has learned from its experiences with new data" (Kopera, 2021).

[8] This matter od adaptive AI was brought to my attention by an anonymous reviewer for which I thank him.

A. Any task that can fulfill system E_1, can also fulfill system E_2. (from the above observations);
B. E_1 will win the game with a human i.e. will accomplish its task. (based on the results from section five);
C. Ergo: E_2 will win the game with a human. (from A. and B.).

This reasoning should show that, despite limiting our considerations to E_1-type systems, we do not limit the resulting conclusions to E_1, but they also apply to E_2.

2. Myths as 'premonitions' of mankind

The existential situation outlined above may—and sometimes does—cause anxiety in some members of the community of human users, as is amply demonstrated in literature. However, it is difficult to point to any facts to which we have access in the form of empirical verification, on the basis of which a rigorous narrative could be constructed regarding the justification of existential anxiety. On the argumentative side, it is difficult to find such premises that would make it possible to justify a conclusion regarding a future threat from AI. Therefore, literary works devoted to this issue are based on speculations and some even treat it as science fiction, while others treat it as a purely logical possibility (cf. article motto). We, however, do not downplay this premonition of humanity (expressed in literary terms), and in this section we will try to explain where this anxiety may come from. Our explanation refers to myths that appeared in human history a long time ago. Mankind, as a species, through its representatives, has created strange stories called *myths*. I call these stories strange because, being created at a very early stage in the development of our species, they speak of problems that have been continuously accompanied it in its

history. For example, we have the myth of the Sphinx, a hybrid winged creature with the body of a lion and the head of a woman, who killed travelers heading to Thebes if they failed to solve a riddle. A similar pattern can be seen in the case of the dragon of Wawel Castle from the legend of Krakus and of many other creatures from various legends. The motif shared by many of these stories is an attempt to defeat or outwit an evil creature, which poses a threat to the normal functioning of the community in which it appears. The origin of this creature is in some sense beyond the natural. Most often, in these stories an attempt to outwit it is a rational act, such as solving the riddle of the Sphinx or giving a dragon a fake sheep filled with sulfur. Some believe—and these are not only so-called *ordinary people* but also distinguished scientists—that in certain situations AI is *per analogiam* appears as the embodiment of a *mythical creature*, which threatens the normal functioning of a society, violating the freedom or privacy of all or some of its members. And, consequently, there is the question of outwitting the creature. Let us repeat that these myths, being a kind of common heritage of mankind, are the real cause of humans' fears and anxieties.

3. Turing's and Searle's tests—AI's first issues

In 1936, the groundbreaking year in the creation of AI,[9] two formal models of computability of effectively computable functions were published: one in the work by the American mathematician Alonzo Church (lambda calculus) and the other in the work by the American

[9] Here we are making a mental shortcut, because strictly speaking it was crucial for AI to create a computer as a machine that practically realises the mathematical idea of computability.

logician—born in Augustów (now in Poland)—Emil Post (machine). A year later the most famous work in the area was published: the one written by the English mathematician and logician, Alan Turing (machine). Very soon Turing's theoretical machines found implementation in the form of real machines, which were called *computers*. This started a discussion about the capabilities of such machines. Turing (1950) proposed what is known today as *the Turing test*, in which a human evaluator judges whether his interlocutor in a conversation held in a natural language is a human or a machine (AI). Turing's intention in this test was to try to answer the question whether a machine can think like a human. He considered the theological argument (one of the arguments against machine thinking), according to which God created man as the only thinking being in the universe, and thinking was a function of the human soul. Without going into the intricacies of the problem here, let us note that according to Catholic theology, thinking belongs to God and to His angels, thus it is not a function of the body or brain as I think Turing probably believed.[10] Turing wanted to convince his contemporaries that a machine can think like a human being in order to contradict underestimation of a machine's capabilities in this regard, widespread at that time. The second test—called the Chinese Room argument—comes from John Searle (1980), and was intended to demonstrate that no digital computer has a kind of "mind" or "consciousness" even if it functionally bears a far-reaching

[10] From a certain point of view, calculation does not intrinsically belong to thinking, because, for example, according to Catholic theology, God, although he thinks, does not calculate, because calculation is the manifestation of a certain kind of ignorance. Cf. for example Isaiah 55:8-9: "For my thoughts are not your thoughts, neither are your ways my ways", declares the Lord. "As the heavens are higher than the earth, so are my [...] thoughts [higher] than your thoughts." Some people believe (basing this belief on the Bible) that God's words in which He speaks of cause and effect justify attributing thinking to Him.

resemblance to human conversational behaviour. Searle's argument dampened the enthusiasm of optimistic proponents of AI and its unlimited possibilities. It can be said that the texts mentioned in this section share a common goal: to take a stand in the argument about machine thinking at a time when, in popular understanding, a machine's skills were not appreciated. From the point of view of this paper, both texts (Turing's and Searle's) have lost a great deal of relevance since their publication, as so much has changed in this area. Summing up what has been said so far briefly and succinctly: today no one asks whether machines can think but rather what machines can do in terms of thinking and intelligence and where the upper limits of their capabilities lie.

4. Preliminary formulation of the fundamental question

Taking into account what has been outlined above, we can say that, from a particular point of view, AI can appear as an element of reality—as an artefact—which poses a threat to society on the way to the unrestricted realization of its development. Therefore, we should look for means by which we could 'outsmart' this *creature*. In different words, we can ask if a human can stand up to AI that controls him.[11] This is another way of phrasing the question posed in the title of the paper.

[11] An extreme case of this is fictionally considered in the plot of the 2009 film "Echelon Conspiracy".

5. Preparatory analysis to adopt a model for consideration

Due to the generality of the problem under consideration and the lack of precision, we are forced to adopt a theoretical model that will at least allow us to answer a part of the question. In the initial paragraph of the paper, we mentioned two senses of interactivity of AI. The second sense, referring to a need for AI to be managed by a human, will not be addressed here, since the anxiety linked with AI concerns the case when AI acquires self-awareness and escapes any human power over itself.[12] Second, we will assume that such AI essentially remains a machine. Subject literature devoted to this area is extensive, especially after the publication in 2015 of the famous open letter "An Open Letter: Research Priorities For Robust And Beneficial Artificial Intelligence", which is now signed by about eight thousands of people involved in science, mostly AI professionals.[13] The authors write there: "[w]e recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial: our AI systems must do what we want them to do [...]". An important work, though long forgotten, is (Good, 1965), in which its author introduces the key concept of *singularity*, understood as the point in human history when an ultraintelligent machine will appear. We think there is one term that frequently appears in subject literature used

[12] My attention has been drawn to a film entitled "Saturn 3" (1980), whose plot considers a similar case.

[13] The letter was signed by world-famous scientists and experts, including Elon Musk and Steven Hawking.

to describe the state of our knowledge on the functioning of AI (or certain algorithms), namely *opacity.*[14] Webster's dictionary gives the following definitions of this term:

- obscurity of sense;
- the quality or state of being mentally obtuse.

This state of opacity affects a large area of the phenomena of social life and raises concerns in some members of the global society. This opacity results in a lack of information and—in a sense—makes the entire area of reality epistemically inaccessible. However, this situation is not unique because similar situations appear in the case of the area of knowledge about the spiritual realm, in particular about God, and also in the microworld studied by quantum mechanics. Although God and the microworld are radically different from the world in which we live, we can sometimes find certain similarities, which we call analogies, between them. It is when we find ourselves in a situation which is analogous but epistemically limited with respect to some reality that we often use a research method called the argument from similarity. Its scheme may look like this:

- An object (of type) X and an object (of type) Y are similar (which is symbolically denoted as $X \approx Y$);
- The similarity follows from P;
- Theorem T holds about object X (symbolically: T(X));
- Therefore: theorem T holds about Y (symbolically: T(Y)).

It is worth noting that, in general, we may be dealing with two areas of reality, one of which, i.e. a source to which object X belongs, is well known to us, while the other, i.e. a target to which object Y belongs,

[14] For example, a recent talk by P. Stacewicz at the *Homo informaticus 8.0* conference explicitly dealt with opacity in the context of AI.

is not well known to us. The reasons for this may vary, and in our case they are opacity or lack of information. Thanks to the similarity we have previously found, we believe that we can cognitively *invade* area Y. More precisely, this means that we can transfer our previously acquired knowledge about X into 'knowledge' about Y.[15]

AREA A	AREA B
1. Object X: game $G(A)$;	1. Object Y: using of a program;
Similarity P:	Similarity P:
2. Player **I**; 3. Player **II**; 4. The game consists of moves and leads to a result. 5. Winning: the play belongs to set A; 6. Theorem T(X) holds: the game $G(A)$ will be won by Player **I**.	2. Algorithm (program); 3. A human; 4. It consists of actions (elementary steps) and leads to an effect. 5. Winning: what the algorithm (machine) "wants" is accomplished; 6. Theorem T(Y) holds: what the algorithm has planned will happen.

Table 2: A summary of the points on which the analogy is based.

Of course, in general, the argument from similarity is not deductive. Its Achilles' heel lies in establishing similarity between objects. In our case, we assume it on the basis of the above sketchy and introductory considerations, while we leave the in-depth investigation of the issue for the future. That is why our assumption and thus our model can be accused of lack of soundness with respect to the phenomenon we model. Our response to this accusation is that our

[15] These issues of similarity and argument are closely related to the theory of analogy, but we will not address them here.

approach nevertheless possesses methodological merit because it is at least an attempt to approach the issue. We will now describe area A in detail, on the basis of the above mentioned similarity.

6. Description of the theoretical model, i.e. Area A

Banach-Mazur games were first described in 1930 and were accompanied by a description of a certain problem by the Polish mathematician from the mathematical school of Lvov, Stanisław Mazur, which is recorded in the "Scottish Book" under the number 43. At present, the game is described in the following way.[16]

We will consider an infinite two-person game with complete information, which we denote by the symbol $G(A)$, where $A \subset \omega^{\omega}$, that is A is a subset of the set of all infinite sequences of natural numbers with zero. The symbol $\omega^{<\omega}$ denotes the set of all finite sequences of natural numbers. The *empty sequence* is denoted by $\langle\rangle$, and the *length* od the finite sequence s by $|s|$. We also have two players: Player **I** and Player **II**, who take turns making *moves*, i.e. choices of natural numbers.

Player **I**:	$x_0 \quad x_1 \quad x_2 \quad \ldots \quad \ldots$
Player **II**:	$y_0 \quad y_1 \quad y_2 \quad \ldots \quad \ldots$

Table 3: A graphic representation of the game.

We use the moves of both players to create one infinite sequence of the form: $z := \langle x_0, y_0, x_1, y_1, \ldots \rangle$, which we call *a play* in *a game* $G(A)$. Definitions of players' strategies play a key role:

[16] The formulations of the given definitions and statements mainly after an excellent exposition given by (Khomskii, 2010) and (Soare, 2016).

Definition 1. Strategy σ for Player **I** is the function:

$$\sigma : \{s \in \omega^{<\omega} : |s| \text{ is even}\} \to \omega.\square$$

Definition 2. Strategy τ for Player **II** is the function:

$$\tau : \{s \in \omega^{<\omega} : |s| \text{ is odd}\} \to \omega.\square$$

Definition 3. Let σ be a strategy for Player **I**, and the sequence $y = \langle y_0, y_1, y_2, \ldots\rangle$ be an infinite sequence of *moves* of Player **II**, then:

$$\sigma^* y = \langle x_0, y_0, x_1, y_1, x_2, y_2, \ldots\rangle,$$

where:
$x_0 = \sigma(\langle\rangle)$;
$x_{i+1} = \sigma(\langle x_0, y_0, x_1, y_1, \ldots, x_i, y_i\rangle)$. $\square$

Definition 4. Let τ be a strategy for Player **II**, and the sequence $x = \langle x_0, x_1, x_2, \ldots\rangle$ be an infinite sequence of *moves* of Player **I**, then:

$$x^* \tau = \langle x_0, y_0, x_1, y_1, x_2, y_2, \ldots\rangle,$$

where:
$y_0 = \tau(\langle x_0\rangle)$;
$y_{i+1} = \tau(\langle x_0, y_0, x_1, y_1, \ldots, x_i, y_i, x_{i+1}\rangle)$. $\square$

If $A \subset \omega^\omega$ is a pay-off set, then:

- Strategy σ is **a winning strategy** for Player **I** in Game $G(A)$ iff for all $y \in \omega^\omega$, we have: $\sigma^* y \in A$.
- Strategy τ is **a winning strategy** for Player **II** in Game $G(A)$ iff for all $x \in \omega^\omega$, we have: $x^* \tau \notin A$.

With the above definitions, we can formulate the axiom of determinacy (**AD**):

(**AD**) For each set $A \subset \omega^\omega$, Game $G(A)$ is determined i.e. exactly one of the players has a winning strategy for $G(A)$.□

This axiom contradicts the axiom of choice in the sense that the axiom of choice implies the existence of an undetermined infinite game.

7. Useful theorems and Theorem T

Theorem 7.1. *Let $A \subset \omega^\omega$ be a countable set, then Player **II** has a winning strategy in the game $G(A)$.* (Khomskii, 2010, p.14)[17]

A slight modification of this concept of the game $G(A)$ is the Banach-Mazur game $G^{**}(A)$, where players alternately choose finite sequences of numbers. In our case, we can treat these two game concepts as equivalent, thanks to the adoption of some coding process of finite sequences.[18] For the second concept we have theorems:

Theorem 7.2. *Player **I** has a winning strategy in the Banach-Mazur game $G^{**}(A)$ iff A is comeager* (Soare, 2016, p.213).

Corollary 7.3. *Player **II** has a winning strategy in the Banach-Mazur game $G^{**}(A)$ iff A is meager* (Soare, 2016, p.213).

The meager and comeager sets relate to Baire space ω^ω. Intuitively "comeager sets are large. They form a filter, are dense, uncountable, and are closed under countable intersections. Meager sets are small. They form an ideal, and countable sets are meager." (Soare, 2016, p.212). I mention this because the matter may be of interest to philosophers.

---

[17] In parentheses I give places from the literature where the proofs of these theorems can be found. When there is no such indication then the claim with the proof comes from me.

[18] I owe my attention to this issue and its clarification to Prof. Yurii Khomskii.

Corollary 7.4. *Let $A = GR_1$, then there is a winning strategy for Player **II**.*

Proof. Let $A = GR_1$, i.e. A is the set of all unary general recursive functions. The set GR_1 is countable. Therefore, by virtue of theorem 7.1, there exists a winning strategy for Player **II**. □

Corollary 7.5. *Let $A \subset \omega^\omega$ and $A = \{f : f \in GR_1$ and $f(n) = s$, for even $n\}$. Then Player **II** has a winning strategy.*

Proof. $|A| < |\omega^\omega|$, then by virtue of theorem 7.1 we have the thesis. □

Corollary 7.6. *Let $A \subset \omega^\omega$ and $A = \{f : f(2n) = s\}$. Then Player **I** has a winning strategy.*

Proof. Since $|A| = |\omega^\omega|$, we cannot use Theorem 7.1. The winning strategy for Player **I** consists in continuously choosing a constant s, i.e, $\sigma(\langle x_0, y_0, x_1, y_1, \ldots, y_i\rangle) := s$, for any i. Let any $y \in \omega^\omega$ be a sequence of the moves of Player **II**. Then for each y, $\sigma^* y = z$, where $z := \langle x_0(= s), y_0, x_1(= s), y_1, \ldots\rangle$. Sequence z has such a form that for each i, $z(2i) = s = x_i = f(2i)$ for any function $f \in A$. □

Theorem T. Let $A \subset \omega^\omega$ and $A = \{f : f|_P \in GR_1$, where $f|_P$ is the restriction of function f to set P of all even numbers$\}$, then Player **I** has a winning strategy.

Proof (sketch). The proof runs along the line of the proof of the previous corollary, except for the fact that for each y, $\sigma^y = z$, where $z := \langle x_0, y_0, x_1, y_1, \ldots, \rangle$. Sequence z has such a form that for each i, $z(2i) = g(i)$ for some fixed function $g \in GR_1$. □

8. Transfer of Theorem T to Area B—T(Y)

In this section we will perform the final step announced in section five, which is the argumentation step we are entitled to by virtue of the argument from similarity presented above. Undoubtedly, this step is quite problematic. We assume that we have established similarity between certain infinite cases of Banach-Mazur games and the use of an algorithm. Again, we omit here the somewhat complicated matter of implementing an algorithm on a machine, and (making a shortcut) we will also talk that a human uses a machine (computer). A machine and a human are understood here as players, where an algorithm (machine) is Player **I** and a human is Player **II**. The moves of a machine consist in giving orders, while a human responds to them by performing some operation on the machine. From a certain point of view, we can look at a machine as a place where a game is played and where an algorithm 'meets' the human mind. We model both types of moves as alternating choices of natural numbers by both players in the form of one infinite sequence of natural numbers. Theorem T formulated above precisely expresses an intuition that for Player **I** it is sufficient to always generate a recursively enumerable recursive sequence in the game. Let us now turn to objects similar in terms of the relation of similarity, i.e. to the counterparts of Player **I** and Player **II**, namely an algorithm and a human. As a result of its action, an algorithm should always generate a sequence which is recursively enumerable. We assume this on the basis of Church's thesis and believe that an algorithm cannot actually generate a sequence other than a recursively enumerable sequence.[19] Hence, the sequence resulting from the game

[19] This passage requires a longer explanation, but due to lack of space, it is not provided here.

will always have such a recursively enumerable set on even positions. This state of affairs will make it possible to accomplish what was coded in the program (algorithm).

Let us now formulate Theorem T(Y), which—for obvious reasons—will not be precise in the case considered in the paper:

Theorem T(Y). Any program (algorithm) implemented on a machine will accomplish the goal written into the program, and an *ordinary user* cannot change it, which means that the user will always lose.

9. Crackers

Unlike Theorem T(X), which is a mathematical theorem that is always true, Theorem T(Y) is empirical, and thus a counter-example can be found for it. A counter-example in such situations can be generated because of an unforeseen gap in the understanding of basic terms. A group of unusual computer users, called *crackers*, generate counter-examples to Theorem T(Y). The unusual nature of these users of a machine, somewhat akin to hackers, is expressed in their setting a goal for themselves to overcome the limitation implied by Theorem T(Y).[20] There are essentially two ways in which crackers operate: breaking into a program and breaking into a server. Crackers are not ordinary computer users, they are often very knowledgeable and competent in certain areas of computer science, and their inexperienced followers are called *script kiddies*. Crackers break firewalls, i.e. these features of a program which are to ensure victory in the game to the algorithm, thus, essentially they break the rules of the game. Banach-Mazur games do not provide for such cases, although from

[20] Note that a Banach-Mazur game does not allow players to break the rules of the game.

the mathematical side this does not change anything because only the assumptions of a theorem become unfulfilled, and the theorem becomes empty satisfied. Thus, in the context of the main question of the paper, we can make this optimistic prediction:

- For any algorithm, if an algorithm follows a program written by a human, then its *cracker* exists.

A pessimistic prediction referring to the notion of *singularity* would sound like this:

- There exists such an algorithm, not necessarily written by a human, that its *cracker* does not exist.

10. Conclusions

Let us finally take a brief look at the entirety of the argument presented in this paper to help the reader grasp its structure. We will list it in points:

i. We began with some people's existential anxiety about AI;
ii. We outlined the definitions of the key terms:
iii. We pointed out the role of myths and legends in the analyzed issue;
iv. We posed the problem;
v. We analyzed the stages of building a model based on similarity;
vi. We described a mathematical model in the form of Banach-Mazur games;
vii. We formulated a fundamental theorem on mathematical games;
viii. We transferred this theorem to the area of computer (algorithm)-human relations in the form of:

a. If an algorithm is correctly defined, a human, as an ordinary user, is unable to prevent it from completing the task written in the program;

ix. We provided a counter-example to the theorem—the cracker problem.

What are the conclusions of the paper? First, the main conclusion is that the computer, understood as an algorithm, will always win in a "confrontation" with an average representative of the human race. Second, based on experience, we know that there are users, specially educated, who are able to outsmart the computer. The third conclusion is that the adequacy of the theoretical model in the form of Banach-Mazur games for the considered problem should be further discussed and this model—as it seems theoretically promising—deserves further investigation.

Bibliography

Artificial Intelligence, 2022. Available at: <https://www.oed.com/viewdictionaryentry/Entry/271625> [visited on 4 January 2023].

Bringsjord, S. and Govindarajulu, N.S., 2022. Artificial Intelligence. In: E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy* [Online]. Fall 2022. Stanford, Calif.: Metaphysics Research Lab, Stanford University. Available at: <https://plato.stanford.edu/archives/fall2022/entries/artificial-intelligence/> [visited on 4 January 2023].

Burns, E., Laskowski, N. and Tucci, L., 2022. *What is Artificial Intelligence (AI)? Definition, Benefits and Use Cases*. Available at: <https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence> [visited on 4 January 2023].

Dodig-Crnkovic, G., 2022. In search of common, information-processing, agency-based framework for anthropogenic, biogenic, and abiotic cognition and intelligence. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (73), pp.17–46.

Dubhashi, D. and Lappin, S., 2017. AI Dangers: Imagined and Real. *Communications of the ACM* [Online], 60(2), pp.43–45. https://doi.org/10.1145/2953876.

Good, I.J., 1965. Speculations Concerning the First Ultraintelligent Machine. In: F.L. Alt and M. Rubinoff, eds. *Advances in Computers* [Online]. Vol. 6. New York: Academic Press, pp.31–88. https://doi.org/10.1016/S0065-2458(08)60418-0.

Khomskii, Y., 2010. *Infinite Games. Summer Course at the University of Sofia, Bulgaria* [Online]. Available at: <https://www.math.uni-hamburg.de/home/khomskii/infinitegames2010/Infinite%20Games%20Sofia.pdf> [visited on 4 January 2023].

Kopera, G., 2021. *How Adaptive AI Outpaces Traditional AI Capabilities*. Available at: <https://www.thoughtai.org/post/how-adaptive-ai-outpaces-traditional-ai-capabilities> [visited on 4 January 2023].

Krzanowski, R. and Polak, P., 2022. The Meta-Ontology of AI systems with Human-Level Intelligence. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (73), pp.197–230.

Russell, S.J. and Norvig, P., 2021. *Artificial Intelligence: A Modern Approach*. Fourth, *Pearson series in artificial intelligence*. Hoboken: Pearson.

Searle, J.R., 1980. Minds, Brains, and Programs. *Behavioral and Brain Sciences* [Online], 3(3), pp.417–424. https://doi.org/10.1017/S0140525X00005756.

Soare, R.I., 2016. *Turing Computability* [Online], *Theory and Applications of Computability*. Berlin; Heidelberg: Springer. https://doi.org/10.1007/978-3-642-31933-4.

Turing, A.M., 1950. Computing Machinery and Intelligence. *Mind* [Online], 59(236), pp.433–460. https://doi.org/10.1093/mind/LIX.236.433.

Perspective on Turing paradigm: An essay

Kazimierz Trzęsicki
University of Białystok, Poland

Abstract

Scientific knowledge is acquired according to some paradigm. Galileo wrote that the "book of nature" was written in mathematical language and could not be understood unless one first understood the language and recognized the characters with which it was written. It is argued that Turing planted the seeds of a new paradigm. According to the Turing Paradigm, the "book of nature" is written in algorithmic language, and science aims to learn how the algorithms change the physical, social, and human universe. Some sources of the Turing Paradigm are pointed out, and a few examples of the application of the Turing Paradigm are discussed.

Keywords

Galileo Galilei, Alan Turing, Konrad Zuse, zero, Arabic numeral, paradigm, mathematics, algorithmics.

> *Tolle numerum omnibus rebus, et omnia pereunt.* [Take from all things their number and all shall perish.]
> (Isidore of Seville, 1911, Liber III, *De mathematica*, IV. *Quid praestent numeri*)

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 281–332

1. Introduction

Science is developed and created according to a pattern, a paradigm. Paradigm is historically mutable. The place of one paradigm is taken by the one which enables a fuller understanding and a better description of the information obtained. The Aristotle paradigm of natural knowledge has been replaced by the paradigm we used to tie with Galileo. According to this paradigm modern science was created. Appointed by Aristotele paradigm of logic lasted until Gottlob Frege. The Greek paradigm of mathematics has been replaced by a paradigm that we can associate with Descartes.

Are the paradigms of modern science not of a historical nature, will further research not lead to new patterns in the practice of science? Information philosophy[1] poses a new paradigm, which I call the Turing paradigm. The Turing paradigm seems to better and more fully capture knowledge in areas where the Galilean paradigm dominates, but also about areas where the Galileo paradigm encounters various limitations.

The key concepts of information philosophy are the concepts of information, algorithm, and artificial intelligence. If we were to briefly characterize the digital age in which we live, three terms would suffice: information, algorithm, artificial intelligence.

---

[1] There is no one definition of information philosophy. It may be characterized, e.g. (Floridi, 2009, p.154): "as the philosophical field concerned with (a) the critical investigation of the conceptual nature and basic principles of information, including its dynamics, utilization, and sciences, and (b) the elaboration and application of information-theoretic and computational methodologies to philosophical problems." See also (Floridi, 2002).

2. Binary notation

The idea of binary code has a long history (Ligonnière, 1992; Trzęsicki, 2006b). Leibniz, creating his binary system, indicated as predecessor the thirteenth-century Arab mathematician Abdallah Beidhawa.

It is usually stated that binary notation was invented and first formally proposed by Leibniz as an illustration of his dualistic philosophy, but already around 1600 the notation was used by an English astronomer Thomas Harriot. John Shirley (1951) writes about his achievements:

> the mathematical papers of Thomas Harriot (1560–1621) show clearly that Harriot not only experimented with number systems, but also understood clearly the theory and practice of binary numeration nearly a century before Leibniz's time.

Several manuscripts of the legacy of Thomas Harriot are evidence that he is probably the first inventor of binary system. He uses 0 and 1 and shows examples how to convert expressions written in the decimal system to expressions written in the binary system and *vice versa*. He demonstrates the basic arithmetic operations, too (Ineichen, 2008). As the first text on the binary system Ineichen points the two-volume work *Mathesis biceps vetus et nova* (1670) by Juan Caramuel y Lobkowitz (Ioannis Caramuelis). In connection with these works by Harriot and Caramuel, the question is raised as to whether Leibniz plagiarized. This question is answered in the affirmative (Ares et al., 2018).

The first binary encoding of alphanumeric characters was done by Giuseppe Peano. In the years 1887–1901 he designed an abstract shorthand machine based on the binary coding of all syllables of the Italian language. Together with the phonemes with 16 bits (so

it had 65,536 combinations), 25 letters of the (Italian) alphabet and 10 numbers were encoded. Peano's code went unnoticed and was forgotten.

The use of binary code was not obvious. Completed in the summer of 1946 American *ENIAC*, unlike binary coded *Z3*, *ABC* and *Colossus*, was based on decimal arithmetic.

The use of the binary system in computers was finally determined by *Burks-Goldstine-Von Neuman Report* to U.S. Army Ordnance Department (finished June 28, 1946) reprinted in (1987, p.105), in which we read:

> An additional point that deserves emphasis is this: An important part of the machine is not arithmetical, but logical in nature. Now logics, being a yes-no system, is fundamentally binary. Therefore, a binary arrangement of the arithmetical organs contributes very significantly towards a more homogeneous machine, which can be better integrated and is more efficient.[2]

3. The world build of numbers

The second fundamental idea for Turing's paradigm is idea of the number as the principle of the world. It has its protagonist in the person of Pythagoras, who proclaimed, as reported in (Guthrie and Fideler, 1987, p.21) that number is the principle, the source and the root of all things. He argued that every existing thing has a numerical value, and in the Middle Ages it was expressed by: *dictum omne ens est*

[2] NB. This explanatory passage was not present in first edition of the report (cf. Burks, Goldstine and von Neumann, 1946, p.13), but was added in later editions (ZFN editor's footnote).

scibile [all beings are knowable](Cherry, 2017, pp.135–136; see also Heschmeyer, 2012). This concept of the number as the principle of the world finds new associations when the idea of zero arises.

In January 1697, Leibniz sent a letter to his protector, Prince Rudolf August of Braunschweig (Herzog von Braunschweig-Wolfenbüttel Rudolph August) with birthday wishes (Leibniz, 1697), in which he discusses the binary system and the idea of creating with 0 as nothingness and 1 as God (Swetz, 2003).

For Leibniz (1697) nothingness and darkness correspond to zero, and the radiant spirit of God corresponds to one. For he believed that all combinations arise out of oneness and nothingness, which is similar to saying that God made everything out of nothing and that there were only two principles: God and nothingness. He designed a medal whose leitmotif was *imago creationis* and *ex nihil ducendis Sufficit Unum*. One is the sun that radiates onto the shapeless earth, zero.

The idea that everything is made of 0 and 1 is the reason why one of the creators of algorithmic information theory, Gregory Chaitin—as he writes not quite seriously—proposes to name the basic unit of information not "bit" but "leibniz" (Chaitin, 2004a; cf. Trzęsicki, 2006a):

> all of information theory derives from Leibniz, for he was the first to emphasize the creative combinatorial potential of the 0 and 1 bit, and how everything can be built up from this one elemental choice, from these two elemental possibilities. So, perhaps not entirely seriously, I should propose changing the name of the unit of information from the bit to the leibniz!

The unit "leibniz" could be the unit (parcel) that Hobbes (1651, Chapter V. Of Reason, and Science) wrote about:

> When a man reasoneth, hee does nothing els but conceive a summe totall, from Addition of parcels.

Leibniz was convinced that the world was organized according to the rules of mathematics. This thought is summarized in the sentence (1890a, p.191)[3]:

> *Cum Deus calculat et cogitationem exercet, fit mundus*.

Mathematics is a tool of the World Constructor, and numbers are the material the world is made of.

Today, the idea of the world as made of mathematical objects, Mathematical Universe Hypothesis, is proclaimed by cosmologist Max Tegmark (2008; 2014). Mathematical objects exist in 'Platonic heaven'. According to Tegmark they are more basic to the universe than atoms and electrons.

4. Modern Science

The idea of the mathematical nature of the world lays at the basis of modern natural science, and its beginning is usually associated with Galileo Galilei, who proclaimed that the book of nature is written in the language of mathematics.

The shaping of the modern paradigm of science in what was then called "natural philosophy" was in fact a revival of the concept of Archimedes (Heller, 2013, pp.71, 77). This idea continued in the Middle Ages. For Roger Bacon (1214–1292) there are four great sciences without which others cannot be known and the meaning of things cannot be understood. And when they are known, then wisdom

[3] More on this entry in the margin of the essay *Dialogus* (Leibniz, 1890a) see (Kopania, 2018).

will be attained without difficulty and labor, not only in the teachings of man, but also in the divine ones. And the possibilities of each of these sciences are revealed not only because of the wisdom itself, but also in relation to the above-mentioned one. In *Opus Majus* (2010, Pars Quarta, Distinctio Prima, Capitulum I) Roger Bacon emphasizes that:

> Of these sciences the gate and key is mathematics, which the saints discovered at the beginning of the world, as I shall show, and which has always been used by all the saints and sages more than all other sciences. Neglect of mathematics works injury to all knowledge, since he who is ignorant of it cannot know the other sciences or the things of this world. And what is worse, men who are thus ignorant are unable to perceive their own ignorance and so do not seek a remedy.

About the place and role of experiments in *De scientia experimentorum: que dicitur dignior Omnibus Partibus Philosophie Naturalis de Perspective: Et ideo notanda est maxime*, a part of *Opus Tertium* (1912), he wrote that the strongest argument proves nothing so long as the conclusions are not verified by experience. Experimental science is the queen of sciences, and the goal of all speculation.

Galileo justifies heliocentrism by referring to the exegesis of *Bible* based on the doctrine of St. Augustine, in particular his *De Genesi ad litteram* (Galileo Galilei, 1615; cf. Sibley, 2013, p.73). In this tradition, which Galileo finds explicitly, the book of nature should be read with mathematical tools rather than those of scholastic philosophy. The book of nature was written in the language of mathematics, and therefore must be interpreted by mathematicians, not theologians. The book of nature as a mathematics contains truths that cannot be discussed.

Galileo (1623, p.4) writes:

> Philosophy is written in this grand book, the universe, which stands continually open to our gaze. But the book cannot be understood unless one first learns to comprehend the language and read the letters in which it is composed. It is written in the language of mathematics, and its characters are triangles, circles, and other geometric figures without which it is humanly impossible to understand a single word of it; without these, one wanders about in a dark labyrinth.

Galileo recommends learning the language of mathematics because it is the language spoken by God (Strogatz, 2019; Wouk, 2010).

Hall (1956, p.97) maintains that:

> Galileo's greatest fame is as an astronomer, yet in intellectual quality and weight his one treatise on mechanics almost outweighs all the rest of his writings. Although his book on cosmology became notorious, and had a more general public influence, it had no comparable effect upon the future development of scientific astronomy, for its polemics were suited only to its own age. The contradiction here is more apparent than genuine. Though formally divided between two branches, Galileo's creative activity in science was a unity, not twofold: it was a unity in laying down revised principles of procedure in science, and again in its specific exemplification of these principles, since Galileo saw that the science of motion and the just appraisal of the results of observational astronomy were the twin keys to an understanding of the universe.

Let us add that Galileo also perceives the mathematical nature of world as its geometricality—such was the Pythagorean tradition. Only Descartes will change this by algebraizing geometry. Descartes' most valuable contribution to the scientific revolution was the co-ordinate geometry (Hall, 1956, p.200). It was only after Descartes that the

thesis proclaimed in *Posterior Analytics* by Aristotle was rejected that arithmetically it was impossible to prove geometric truths. After Descartes, mathematics—which was important for the development of science—became knowledge about functions and operations, not just about numbers.

As a sickly child, Descartes had the privilege of getting up late. He retained this practice as an adult. The German philosopher Daniel Lipstropius took advantage of this and came up with a story of how Descartes got the idea of what we call the Cartesian coordinate system (Mazur, 2014, pp.111–112). Descartes, according to this fairy tale, had this wonderful revolutionary idea for mathematics to come across flies crawling on the ceiling in his bedroom in La Flèche in 1636. He noticed that the position of a fly could be clearly defined by its distance from the walls.

Newton creates calculus because it is the language with which the book of nature is written. Also Leibniz creates calculus. As an aside, let us add that Newton accused him of plagiarism (Sonar, 2018). It just so happens that Leibniz, the genius of creating symbols ('The Symbol Master', cf. Mazur, 2014, pp.165–168)—having a greater understanding of the choice of language—gave his version a linguistic representation that resulted in the development of which failed with the approach proposed by Newton. Charles Babbage, the creator of the first (mechanical) programmable computer, noticing the delay of English mathematics in relation to French mathematics, undertook to translate French texts from mathematics (Trzęsicki, 2006c). Babbage (1864; 2008) wrote:

> Under these circumstances it was not surprising that I should perceive and be penetrated with the superior power of the notation of Leibniz.

For Isaac Newton and other philosophers of this period, the mathematical expression of philosophical concepts also encompassed natural human relationships: the same laws moved physical and spiritual reality. Mathematical models were indicated for human behavior. In the case of Pascal, for example, this is a famous wager: a rational person should live as if God existed. If God does not exist, that person has finite losses (some pleasure, luxury, etc.), and if he exists, he can gain infinitely much (infinite happy life in heaven) and avoid infinite losses (eternity in hell). The wager is the first example of a formal use of decision theory.

Gottfried Leibniz (1697; 1979) mathematically models the creation and composition of the world (Trzęsicki, 2006b,c; 2020a). Following Hobbes, he preaches the concept of thinking as calculus: *cogitatio est calculatio* (Leibniz, 1666). All of this is consistent with the concept of God as the one who creates the world by calculating. Mathematics is a tool of the constructor of the world and numbers are the material from which the world is made. It is a God whose logic is the same as that of man.

According to Johannes Kepler, angels also move planets according to a mathematical model (Donahue, 1993; Wolfson, 1962).

Later the idea of God (God of Spinoza) as a "mathematician" is proclaimed by Einstein (Infeld, 1980, p.279):

> God does not care about our mathematical difficulties. He integrates empirically.

This is according to Heller (2014, p.41):

> A fundamental hypothesis, tacitly taken in the very method of modern mathematized empirical science [which] states that there is nothing in the material world that cannot be mathematically described.

The broadening of the idea of the mathematical nature of world to other fields was proclaimed by many, e.g., Nicolas de Condorcet wrote in *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix* (1785) (Essay on the Application of Analysis to the Probability of Majority Decisions), about the application of calculus to social and political sciences. Politics would then become rational.

In the period before the scientific revolution, it was assumed that nature is rational, because God, its creator, is rational. After the revolution, the rationality of nature was discovered within itself. Studying the natural world is no longer getting to know God. Nature is a mechanism. Kepler wrote that *caelestic machina* was not *instar divini animalis, sed instar horologii* and that Galileo often spoke similarly, especially in his famous adagio *universum horologium est*, the universe is a clock. Descartes thought that animals were merely 'mechanisms' or 'automata' and that as a result, they were the same type of thing as less complex machines like cuckoo clocks or watches.

God is conceived as an engineer. He would be a bad engineer, and he is not if he constantly engages in the operation of this mechanism. Ultimately, it becomes redundant. Pierre Simon de Laplace introduced Napoleon to *Systeme du Monde*. He asked him, "Have you written a huge book on the world system without any mention of the Creator of the universe?" Laplace replied: Sir, I didn't need any such hypothesis. ("*Je n'avais pas besoin de cette hypothése-lá*.") Napoleon told Joseph-Louis Lagrange about it, who exclaimed, "Ah! c'est une belle hypothêse; ça explique beaucoup de choses" (De Morgan, 1872, pp.249–250).

5. Idea of algorithm

The idea of algorithm permeates all sciences, and the beauty of the algorithmic approach correlates with ease of understanding.

Various algorithms were used even before our era. Babylonian mathematicians as early as around 2500 B.C.E., and Egyptian mathematicians around 1500 B.C.E. they calculated the quotient algorithmically. Greek mathematicians used Eratosthenes' sieve to find prime numbers and the Euclid's algorithm to find the greatest common divisor. In 9th century Arabia, cryptographic algorithms were used for decryption.

The name "algorithm" derives from the name of born in present-day Uzbekistan the Persian mathematician Abu Abdullah Muhammad ibn Musa al-Khwarizmi, or rather his Latinized version of "al-Khwarizmi" (Knuth, 1997, p.1). The Latin "algorithmus" is a combination of "algorism" and the Greek "arithmós" (number) (Marciszewski, 1981, p.14). "Algorithmus" (algorismus) meant performing arithmetic operations on numbers written in Arabic numerals, as opposed to performing these operations on numbers written in Roman numerals.

Robert of Chester, who was the first translator into Latin of the now-lost book al-Khwarizmi (Menninger, 1969, p.411), his translation—found in the 19th century—begins with the words:

> *Dixit Algoritmi: laudes deo rectori nostro atque defensori dicmus dignas.*

Around 1143 (Menninger, 1958, p.411) an abstract was made of this work today known as *Salem Codex* (Cantor, 1865). At the beginning we read:

> *Incipit liber algorizmi: omnis sapientia sive scientia a domine Deo; sicut scriptum est: Hoc quod continent omnia scientiam habet, et iterum: Omnia in mensura et pondere et numero constituisti.*

The grammatical form used proves that the author was not aware that it was a surname (Cantor, 1865, p.14, footnote 1). In this text, the name "algorizmi" was for the first time—in the preserved literature—used to mark a procedure:

> Der Gebrauch des Nominativus *algorizmus* beweist, dass das Bewusstsein, dass *Algorizmus* der Name eines Mannes sei, bei dem Verfasser der Abhandlung schon verloren gegangen war. Er hielt offenbar dieses Wort für den Namen der Rechenkunst selbst.

I wonder if the author of *The Code of Salem* refers to the all-embracing knowledge of God to make it sublime, because that was the custom, or if he has any sense of the role and place of algorithms in the work of creation. If the latter, then it can be indicated as the one who anticipated the basic idea of information philosophy, that is, that algorithms guide the events of the world.

In a Latin poem written for didactic purposes and attributed to Alexander de Villa Dei (Alexander de Villedieu) *Carmen de Algorismo* or *Algorismus metricus* (printed edition 1839, p.73) we read:

> Hinc incipit algorismus.
> Haec algorismus ars praesens dicitur in qua
> Talibus Indorum fruimur bis quinque figuris
> 0 9 8 7 6 5 4 3 2 1,

Algorithms are related to the way information is encoded. In other words, a change in the coding method may involve a change in the algorithm. It may be that this change is radical—as one might

suppose—as it is in the case of physical algorithms that process biological code. Let us add after Marciszewski (2011, p.199–200) that by physical algorithms we understand algorithms that control information processing that takes place in physical reality—those that process information that makes up the world—unlike symbolic algorithms that we write and use whose computers process information encoded by us. Natural algorithms perform natural computing. Cognitive calculations, the ones we do, are carried out with the help of symbolic algorithms.

Algorithms should not only—which is obvious—be correct, that is, give a true output, but also, as Donald Knuth writes (1997, p.7):

> we want good algorithms in some loosely defined aesthetic sense. One criterion [...] is the length of time taken to perform the algorithm [...] Other criteria are adaptability of the algorithm to computers, its simplicity and elegance, etc.

Gregory Chaitin (2004b, p.27) specifies the concept of elegance in the program:

> a program is 'elegant', by which I mean that it's the smallest possible program for producing the output that it does.

At the same time, he adds that:

> I'll show you can't prove that a program is 'elegant'—such a proof would solve the Halting problem.

The beauty of natural algorithms and their accessibility to the human mind is inherited by symbolic algorithms.

The definition of algorithm is the work of 20th century mathematicians and logicians. The need for such a definition emerged in connection with Hilbert's program, who postulated the creation of mathematics by formal transformations of the symbolic representation of mathematical knowledge. These transformations were to

be such that there was no dispute as to their correct implementation. Moreover, as is clear, they were to lead from true mathematical sentences to true mathematical propositions, that is, in this way, a possible contradiction was to be ruled out if the original data were not contradictory. Belonging a sentence to a set of statements was to be formally resolved. Such an approach required the definition of the notion of a formal method that would be a tool for the implementation of such an undertaking. Among the proposals—which turned out to be equivalent—the concept developed by Alan Turing, known today as the Turing machine, was particularly appreciated. An algorithm is a procedure that is executable with a Turing machine.

Although the concept of an algorithm defined in this way has been successful, it does not mean that—including Turing (1950)—have ceased to consider modifying the concept of an algorithm.

Let us add that the word "computer" was still in the 19th century, and even in 1936—when Turing published *On Computable Numbers, with an Application to the Entscheidungsproblem* (1937)—used to indicate an official who was doing cumbersome numerical calculations (Copeland, Shagrir and Sprevak, 2016, p.446). Thus understood, "computer" would denote a reckoner, in Polish: "rachmistrz". Polish texts in which "computer" is translated into "komputer" are devoid of any associations present in English texts. In particular, the association of "komputer" with "rachmistrz" is important for the correct understanding of the Turing texts.

6. Reality and information

Information is the content of our knowledge.[4] According to Luciano Floridi (2008), the pioneer of the philosophy of information, reality in itself is not a source but a resource for knowledge. Stehphen Wolfram (2002, p.389) states:

> [M]atter is merely our way of representing to ourselves things that are in fact some pattern of information, but we can also say that matter is the primary thing and that information is our representation of that. It makes little difference, I don't think there's a big distinction—if one is right that there's an ultimate model for the representation of universe in terms of computation.

The retrieved information must be represented somehow. Representation enables it to be stored, communicated and processed. Each piece of information may be zero-one encoded. The way of representation is subordinated to the purpose and what it is supposed to serve. As John Wheeler (1989) puts it:

> every physical quantity, every it, derives its ultimate significance from bits, binary yes-or-no indications.

This idea can be summed up in short: it from bit, where "it" is what exists and "bit" refers to information.

Konrad Zuse (2012b, p.5) developing the concept of digitized spatial relations, the idea of understanding the universe as a computer, assigns an important role to the concept of information:

[4] Stacewicz (2011, §1) excellently discusses the concept of information and its relationship with knowledge.

> In current expanded usage, the term "compute" is identical with "information processing." By analogy, the terms "computer" and "information-processing machine" may be taken as identical.

Zuse was the first to suggest that the physical states of the universe are computed by the universe itself. He pointed to cellular automata. The concept of cellular automata was developed by John von Neumann in connection with his search for similarities between computers and the central nervous system (von Neumann, 1958; 1963; 2012; von Neumann and Burks, 1966; Shannon, 1958).

The information can be processed algorithmically. Aristotle, creating a syllogistic, constructs what today is recognized as a formal information processing system. This idea is developed in formal logic.

Usually, Gottfried Leibniz is mentioned as the one who emphasized and associated the development of knowledge with the applications of computational information processing.

If thinking is a calculation, and the world is made of numbers, we will arrive at all the truth that we can arrive at by calculating. Thus:

> *Quo facto, quando orientur controversiae, non magis disputatione opus erit inter duos philosophos, quam inter duos Computistas. Sufficiet enim calamos in manus sumere sedereque ad abacos, et sibi mutuo (accito si placet amico) dicere: calculemus.* (Leibniz, 1890b, p.200)[5]

The ontological thesis about the world as created by 1 with 0 has opened up new perspectives for combining the concept of information with metaphysics. In praising his binary arithmetic, Leibniz (1990) said:

---

[5] Similar statements can be found in other texts of the cited volume, for example on pages: 26, 64-65, 125.

> *tamen ubi Arithmeticam meam Binariam excogitavi, antequam Fohianorum characterum in mentem venirent, pulcherrimam in ea latereis judicavi ex nuhinem origin reisavi potentiam summae Unitatis, seu Dei.*

This idea fascinated Leibniz so much that he passed it on to Father Grimaldi, a mathematician at the court of the Emperor of China, in the hope that with it he would convert the Emperor and with him to christianize of all China (Leibniz, 1697).

Calculating is an activity in which a machine can replace a human being. In 1685, when discussing the value for astronomers of a calculating machine he invented in 1673, more efficient than Pascal and performing all basic arithmetic operations, Leibniz (1929, p.181) wrote that (Davis, 2001, Ch. I: Leibniz's Dream):

> For it is unworthy of excellent men to lose hours like slaves in the labor of calculation which could safely be relegated to anyone else if the machine were used.

Charles Babbage, when he and his friend were preparing math tables, noticing a lot of errors, was frustrated and shouted (Swade, 2002):

> I wish to God these calculations had been executed by steam!

Konrad Zuse in an interview with Uta Merzbach in 1978 said that when he had to do tedious engineering calculations, the think[6]:

> It's beneath a man. That should be accomplished with machines.

motivated him to understand the work of building a computer (Copeland, Shagrir and Sprevak, 2016, p.449).

---

[6] Konrad Zuse interviewed by Uta Merzbach in 1968 (Computer Oral History Collection, Archives Center, National Museum of American History, Washington DC).

This pragmatic argument with the above-mentioned metaphysical arguments inspired computer scientists and motivate their aims towards creating of artificial intelligence. If all truth has a numerical representation, and thinking is represented by numerical operations, all of which can be done by a calculating machine.

The idea of mechanical acquisition of knowledge, *ars combinatoria*, having ancient roots, and in Europe propagated and developed by Lullists, i.e. those who referred to the concept of Raymondus Lullus (Trzęsicki, 2020a,b), had to be popular in the 17th century, if we also find literary references to it. Jonathan Swift, an Irishman, twenty-one years younger than Leibniz, in 1726 in *Gulliver's Travels* (1892) literally illustrates this idea:

> The first professor I saw, was in a very large room, with forty pupils about him. After salutation, observing me to look earnestly upon a frame, which took up the greatest part of both the length and breadth of the room, he said, "Perhaps I might wonder to see him employed in a project for improving speculative knowledge, by practical and mechanical operations. But the world would soon be sensible of its usefulness; and he flattered himself, that a more noble, exalted thought never sprang in any other man's head. Every one knew how laborious the usual method is of attaining to arts and sciences; whereas, by his contrivance, the most ignorant person, at a reasonable charge, and with a little bodily labor, might write books in philosophy, poetry, politics, laws, mathematics, and theology, without the least assistance from genius or study."

7. The concept of a paradigm and its implementations

The term "paradigm" is derived from Greek: παράδειγμα (parádeigma) which translates to "example", "pattern", "template" or "explanation model", "seeing the world", "worldview". The term "paradigm" was popularized by Thomas Kuhn in the book *The Structure of Scientific Revolutions* (1962; 1974).[7] However, the term was already used by Plato in *Timaeus* to designate a model, the pattern that Demiurge used to create the cosmos.

The paradigm includes philosophical and methodological assumptions commonly and permanently adopted by those who practice science at some stage of its development. Knowledge is divided into paradigmatic, that is, scientific, and pre-paradigmatic, that is, pre-scientific.

A paradigm is a pattern for doing science at some stage of its development. The new pattern, the new paradigm, dismisses as (already) unscientific some of the problems of the old science, and gives new meaning to those that remain in the new science. Moreover, importantly, it solves problems that science could not cope with in the previous version of the paradigm and sets new questions.

Galileo proclaimed—which led to the designation of a paradigm of science different from the Aristotelian one—that the book of nature is written in the language of mathematics, and therefore this language is appropriate for knowing and understanding it. Mathematical natural science is practiced according to the Galileo paradigm.

Note that in Galileo's time the state of mathematical knowledge was far from what it is today. The mathematics of Galileo's day are

[7] Kuhn's idea of paradigm has been the subject of discussion, criticism and modifications.

different from that of science today. The development of mathematics was coupled with the progress of natural science. For example, Newton creates calculus for the sake of his "natural philosophy".

Creating science according to the Galileo paradigm not only resulted in a deeper understanding of the natural world, but also brought fruit in the form of technology, which led to the development of industry, as well as changes in social relations (Marciszewski and Stacewicz, 2011, p.141–148).

The Turing paradigm will be understand as a paradigm that assumes that the book of natural reality is written in some universal programming language and that this language is the proper language of knowledge about both natural phenomena and about any other cognitively available to man in the natural order. The paradigm is build on the legacy of Turing's computationalism, the view that nature physically computes its own time development. The idea of such a new paradigm was stated by Konrad Zuse. In his autobiography we read (Zuse, 2012b, p.63–64):

> In the final analysis, the concept of the computing universe requires a rethinking of ideas, for which physicists are not yet prepared. Yet it is clear that earlier concepts have reached the limits of their possibilities; but no one dares to switch to a fundamentally new track. Yet, with quantization, the preliminary steps towards a digitalization of physics have already been taken; but only a few physicists have attempted to think along the lines of these new categories of computer science. [...] This was illustrated quite clearly during the conference on the Physics of Computation, held May 6–8, 1981 [at MIT]. What was typical at this conference was that, although the relationship between physics and computer science, and/or computer hardware, was examined in detail, the questions of the physical possibilities and limits of computer hardware still dominated the discussions. The deeper question, to what extent

> processes in physics can be explained as computer processes, was dealt with only marginally at this otherwise very advanced conference.

The Turing paradigm is not in opposition to the Galileo paradigm, but rather clarifies and modifies it. However, it has paradigm-specific consequences, overruling certain problems primarily in the areas of biology, psychology, and sociology, and opens up perspectives for research that—speaking freely—was not visible or not so visible from the perspective of the Galileo paradigm, such as the issue of mind, social and economic life (Marciszewski and Stacewicz, 2011, chapter 20).

Gaston Bachelard (2002) introduced the concepts of epistemological obstacle and epistemological break (*obstacle épistémologique* and *rupture épistémologique*). Science does not progress uniformly linearly. An epistemological break—the term popularized by Louis Althusser—occurs when the integration of the old theory into the new paradigm takes place.

Darwin's evolutionary paradigm appears to be incompatible with the Galilean paradigm, while composing and complementing each other with the Turing paradigm. Computing is more than a language of nature as computation produces real time physical behaviors. The Turing paradigm covers not only natural science, but everything that has traditionally been called natural philosophy. It enables comprehensive research of self-organizing adaptive systems, regardless of their type (physical, biological, social) (Dodig-Crnkovic, 2013; 2022).

8. The world created by algorithms

The concept of algorithm is fundamental to the Turing paradigm.[8] This does not mean that the concept is ultimately defined and closed to changes and modifications. Like the mathematics of the Galileo paradigm, it is alive and coupled with the development of research. Marciszewski writes (Marciszewski and Stacewicz, 2011, p.164) that the intellectual intuition and ingenuity of the scientist are what enable the emergence of new algorithms that will strengthen the computer science system so much that problems in the previous the undecidable phase will become possible to be resolved in an algorithmic manner. In the new system, new undecidable problems will arise, but there is again a chance to overcome difficulties thanks to creative intuition. It turns out that the process of learning about the mathematical world with the use of machines is never closed in the sense of having final results, but is never closed in the sense of the impossibility of further development. It is possible to develop endlessly.

Turing not only gave a definition of an algorithm, a Turing machine, but also indicated new areas of adapting the concept of an algorithm to research needs.

Turing—at least among those with a background in algorithmic science—was the first to embrace the idea of what we call Turing paradigm.

Computing Machinery and Intelligence (1950) can still be a source of inspiration in creating and developing the algorithmic paradigm. Alan Turing, ending his considerations in *Computing Machinery and Intelligence* (1950, p.64) notices the inconvenience of a systematic solution method and writes:

---

[8] For this reason the Turing paradigm may be referred as "algorithmic paradigm" or as "computer science paradigm".

> We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc. Again I do not know what the right answer is, but I think both approaches should be tried.

In this heroic phase of the history of computer science, as Marciszewski (2011, p.165) claims—in addition to Turing, an important contribution is made by John von Neumann, who laid the foundations for the algorithmic paradigm. Von Neumann went further—albeit in the same way as Turing—in postulating an understanding of the algorithm. In his unfinished book, *The Computer and the Brain* (1958), written before his death, he examines algorithms whose carrier would be a living protein.

Considering the possibilities of artificial intelligence, which could be displayed by a machine built according to the rules and principles of the mechanistic paradigm, lead to the conclusion that as such it will not be equal to the intelligence displayed by living organisms (Trzęsicki, 2016).

Konrad Zuse[9] also belongs to this heroic phase of history. He was a pioneer of computer science, although his name is less widely known. Zuse built the first fully programmable $Z3$ computer in the 1940s. The *Plankalkül* programming language was ahead of what others came later. Let us add that the scale and value of Zuse's technical achievement is under discussion (Copeland, Shagrir and Sprevak, 2016, p.448).

[9] http://www.konrad-zuse.de.

There are many parallels between Zuse's and Turing's interests (Zuse, 2012b, p.58). Though Zuse and Turing never met but they became acquainted with each other's work (Zenil, 2012, p.60).

Zuse in *Rechnender Raum* (1967) is the first to talk about the universe as a computer network. He does not announce that he has a complete theory of everything in the form of some algorithm for counting the universe, but in this text he is the first to clearly formulate such an idea. He published the results of his further reflections in *Rechnender Raum* (1969) translated as *Calculating Space* (2012a). In *Der Computer* (2010) he mentions:

> Es geschah bei den Betrachtungen über die Kausalität, daß mir plötzlich der Gedanke auftauchte, den Kosmos als eine gigantische Rechenmaschine aufzufassen. Ich dachte dabei an die Relaisrechner: Relaisrechner enthalten Relaisketten. Stößt man ein Relais an, so pflanzt sich dieser Impuls durch die ganze Kette fort. So müßte sich auch ein Lichtquant fortpflanzen, ging es mir durch den Kopf. Der Gedanke setzte sich fest; ich habe ihn im Laufe der Jahre zur Idee des "Rechnenden Raumes" ausgebaut. Es sollte freilich dreißig Jahre dauern, ehe mir eine erste konkrete Formulierung der Idee gelang.

It was only in the third millennium that the idea of the world as a computer began to attract more attention. Among others, in *Scientific American* and *Spectrum der Wissenschaft* texts such as "Is the universe a Big Computer?", "Is the Universe a Computer?" are published. In the Autumn of 2006 the Technische Universität Berlin, where Horst Zuse, the son of Konrad Zuse then was a professor, organized a conference *Ist das Universum ein Computer?* (Is the Universe a Computer?) (Zenil, 2012, p.61).

Zuse (2012b, p.56), ending with *Rechnender Raum* (1967, p.344), lists the paradigmatic differences between classical mechanics, quantum mechanics and his concept of Rechnender Raum:

lp.	Classical physics	Quantum mechanics	Calculating space
1.	Point mechanics	Wave mechanics	Automaton theory Counter algebra
2.	Particles	Wave-particle	Counter state, digital particle
3.	Analog	Hybrid	Digital
4.	Analysis	Differential equations	Difference equations and logical operations
5.	All values continuous	A number of values quantized	All values have quantized
6.	No limiting values	With the exception of the speed of light no limiting values	Minimum and maximum values for every values for every
7.	Infinitely accurate	Probability relation	Limits on calculation accuracy
8.	Causality in both time directions	Only static causality, division into probabilities	Causality only in the positive time direction introduction of probability terms possible, but not necessary
9.		Classical mechanics is statistically approximated	Are the limits of probability of quantum physics explainable with determinate space structures?
10.		Based on formulas	Based on counters

While Konrad Zuse's concept of the world as a great computer is debatable (Copeland, Shagrir and Sprevak, 2016, paragraphs: "Zuse thesis", and "Examining Zuse's thesis"), the paradigm differences are interesting from the point of view of information philosophy.

Alan Turing did not limit his thinking to computer science issues. He wasn't only looking for knowledge about the mind. His research

also covered the natural world. Not without reason he can be qualified as the philosopher of nature (Hodges, 1997). An example of research according to the paradigm, which we refer to here as the Turing paradigm, is the research, the results of which are included in *The Chemical Basis of Morphogenesis* (1952).

According to mechanicism, everything that is and is happening in nature can be explained by the concepts and laws of mechanics, possibly quantum mechanics. According to information technology, everything that is the subject of scientific knowledge can be explained as algorithmic information processing, analogous to the operation of a Turing machine and its modifications and generalizations, i.e. with the help of the concepts and laws of algorithmics. Computing in the universe, natural computing, would be performed on many different levels of the organization: quantum, biological, spatial, etc. Some computations would be discrete, some continuous (Lesne, 2007).

The difference between the Galilean paradigm and the Turing paradigm, can be shown figuratively: in the concept of science in the Galileo paradigm, the world is the work of a Mechanical Engineer, and in the concept of science in the Turing paradigm, the world is the work of a Programmer. If *Deus ex machina* can be associated with the Galileo paradigm, then with the Turing paradigm we would associate the phrase: *Deus ex AI*.

Accurately, taking into account the historical context, the Turing paradigm can be characterized in the words of Marciszewski (Marciszewski and Stacewicz, 2011, p.153), who in place of Leibniz's statement: *cum Deus calculat et cogitationem exercet, fit mundus* puts: *cum mundus calculat, fit mundus*, when the world counts, the world becomes. Or perhaps, keeping the analogy—bearing in mind the translation "*cum Deus calculat et cogitationem exercet, fit mundus*" as "when God counts and incorporates his thoughts into deeds, the

world arises"—say: *cum mundus calculat et algorithmum exercet, fit mundus*? "*Cum mundus calculat et algorithmum exercet, fit mundus*", translating as "when the world calculates and executes algorithms, the world becomes". The world of Galileo has been calculated, and the world of Turing on the basis of its current state calculates its future state (Chaitin, 2007, p.13).

In the discussion, Marciszewski asks whether the phrases "*calculat*" and "*algorithmum exercet*" are synonymous in the Turing paradigm, or at least equal in scope. If so, the ratio between them would have to be expressed not with "*et*" but (e.g.) "*id est*". And he ponders what Leibniz means when he adds "*et cogitationes exercet*" after "*calculat*". Did he not think that God's thinking is equated with counting? Then what would this add-on be for? Let us add that the last problem posed by Marciszewski is the subject of many considerations, and Louis Couturat (1901) as the motto *La logique de Leibniz*, his fundamental text on Leibniz's logic, has just chosen a shortened form: *cum Deus calculat . . . fit mundus*. If we were to stick to the abbreviated version of Leibniz's thought, then "cum mundus algorithmum exercet, fit mundus" would directly express the idea of information philosophy.

In the world of Galileo, there is an eternal movement defined by the laws of mechanics. The world itself remains eternal and unchanging (steady-state model). The world, however, is not eternal: it has a beginning and will have an end. It begun with the Big Bang and will terminate as Nothing, as the sum of positive and negative energies that are equal one another. The world is not immutable: it evolves. Darwin showed the evolution of the living world. Modern physics states historicity, the evolution of the material world. History teaches about the evolution of the social world.

The laws of mechanics say that the wheels, cogs, and other parts of the machine move, not that the machine itself changes. It's just immersed in the space-time world. The laws of algorithmics speak of processing not only parts, components of the world, but also (the whole) world.

9. The Turing paradigm in science

Let us discuss some scientific questions that are considered differently in the Turing paradigm and in the Galileo paradigm.

9.1 Is the Turing paradigm fruitful?

Does changing the language of mathematics to the language of algorithmics lead to new questions and make it possible to find answers to questions that are not answered in the Galileo paradigm?

In the case of the Galileo paradigm, the natural question is who is calculating. In the case of the Turing paradigm, the answer to the question of what processes information that makes up the world is simple: the world. An algorithm is part of the world just as data and programs are part of a computer, and just as data and programs are both encoded. Let us repeat the sentence which expresses this:

cum mundus calculat et algorithmum exercet, fit mundus.

The cognitive fruitfulness of the Turing paradigm may also—which sounds paradoxical—manifest itself in the statement that some natural and mental processes are not computable. Turing himself, bearing in mind the existence of incalculable real numbers, pointed to the possibility that the physics of the brain may not be computable and

allowed for the possibility of incalculable physical systems (Copeland, Shagrir and Sprevak, 2016, paragraph: "The physical computability thesis").

The practical fruitfulness of the Turing paradigm manifests itself in replacing mechanical technologies with information technologies. There is progress in civilization, as mentioned by Alfred Whitehead (1911, p.61):

> Civilization advances by extending the number of important operations which can be perform without thinking about them. Operations of thought are like cavalry charges in a battle—they are strictly limited in number, they require fresh horses, and must only be made at decisive moments.

Civilization understood in this way will be realized through the development of artificial intelligence, which becomes a "cavalry charge" of the modern world.

In the *Conclusions* of *Rechnender Raum (Calculating Space)* Zuse writes:

> Even if these observations do not result in new, easily understood solutions, it may still be demonstrated that the methods suggested have opened several new perspectives which are worthy of being pursued. Incorporation of the concepts of information and the automaton theory in physical observations will become even more critical, as even more use is made of whole numbers, discrete states and the like.

Stanisław Krajewski (2012) has multiple cognitive hopes with what we call the Turing paradigm. He maintains that due to the advent of computers philosophy has entered a new condition:

> I wish to point out something more fundamental—a new kind of experience with which we have familiarized ourselves because of computers. Much more has happened than the obvious, though still remarkable, 'shrinking' of the globe due to

> the ease of communication with nearly every spot on earth; even more amazing is the fact that so much can be recreated or simulated by programming. The philosophy of mind has been deeply affected by this: indeed, a cognitive science has arisen that conceives the mind as a biological computer. To understand it, the knowledge of logic should be useful. After all, logic, which emerged as the result of an analysis of thinking and thought patterns, was used to build computers, and computers, in turn, according to their enthusiasts, are about to acquire the ability to think. If so, then, however "artificial" this thinking could be, it would amount to not only information processing but to understanding as well.

9.2 Knowing the mind

The Turing paradigm is appropriate and fruitful in the study of the mind. In this area of knowledge, the Turing paradigm is most successful, so that it even seems to be its core field of application at least in the technology and in accomplishment of artificial intelligence.

The problem of the mind in the perspective of the computer science paradigm was taken up by Turing in connection with the death in 1930 of his school friend Christopher Morcom. In 1932, while visiting Morcom's family home, he expressed the conviction, inspired by Arthur S. Eddington's book *The Nature of The Physical World* (2014), that the brain is not deterministic and that free will is based on laws of quantum physics. The result of his reflections is also a test, known today as the Turing test, which prompted the algorithmic understanding of the mind and consciousness. This test became a model for others who set themselves the goal of fully identifying the mind (Krajewski, 2012). Zuse his universal language *Plankalkül* compared to an "artificial brain" (Zenil, 2012, p.62). A new

multidisciplinary science, cognitive science, has become a field of extensive collaboration between researchers of various aspects of the mind and brain.

There are significant achievements in knowledge of the mind. They provide arguments for a negative answer to the title question posed by Włodzisław Duch (2017): "Why Minds cannot be Received, but are Created by Brains". Life after death is supposed to be a myth (Martin and Augustine, 2015). Professor Duch asks: "Will the son of man find faith [...] in information civilization?" Takes up the topic *Catholicism after cognitive science: For a new theology of mind* (Duch, 2015; see also Duch, 2012).

Consider the theological foundations of this discussion of the spirit-body relationship. Does theology really say what is assumed in this discussion? Let us note that the assumption about the separation of soul and body is not an indispensable thesis of Catholic theology. Bocheński (1994) writes that the idea that a man is composed of two pieces, a body and a soul, is a very miserable superstition. All our science and all serious thinkers reject it vehemently. To give just one example, St. Thomas Aquinas, one of the greatest thinkers of Christianity, emphatically denies that the human soul is a "complete substance", that is, a piece, and defends the view that it is "the content (form) of the body."

Does the Thomistic approach to the relationship of soul and body fit into the computer science paradigm? We do not intend to answer these and other questions here, but note that Christians preach a resurrection with body and soul, that the end of this world is not the end of the world at all. As we read in Revelation (21: 1–2):

> I saw a new heaven and a new earth. The first heaven and the first earth had disappeared, and so had the sea. Then I saw New

> Jerusalem, that holy city, coming down from God in heaven. It was like a bride dressed in her wedding gown and ready to meet her husband.

The end of the world would be if all (physical) algorithms were to stop working, towards improvement because the world would be perfect.

Materialistic philosophy accepts the concept of the mind as an exclusively material object. Lenin's brain, who died in 1924, was dissected and tested in a dedicated institute. The aim was to obtain biological knowledge about the genius's brain, and by preserving Lenin's body, it was allowed to revive it. This approach took place in the Galileo paradigm. This biological concept of research essentially narrows down the methods in relation to the Turing paradigm.

The Turing paradigm opens science to speculations about the mind that are made on the Gödel theorem and its versions leading to the rejection of the mechanistic concept of mind (Krajewski, 2020).

9.3 Prediction

We practice science in order to be able to make predictions. As the philosopher of positivism August Comte put it:

> *Savoir pour prévoir, prévoir pour pouvoir.*

In the mechanistic paradigm, the inadequacy of prediction is explained by the scarcity of relevant data or—possibly—of insufficient knowledge about the laws governing the reality under consideration.

The mechanistic paradigm is successful in the field of macro-natural phenomena: we are able to predict the movements of celestial bodies with an accuracy limited only by the errors of observation instruments. It's a bit worse at the micro level, but it works. When, however, social phenomena, e.g. economic, are predicted according to this paradigm, then even simplifying management—as was done

in a centrally planned economy—by administering prices, production volumes, distribution rules and other elements affecting economic performance, you experience a lack of predictability. Why? Perhaps simply because a mechanistic model of the functioning of the economy was assumed. The troubles of the Soviet economy motivate Victor Glushkov idea of OGAS (ОГАС, Общегосударственная автоматизированная система учёта и обработки информации, National Automated System for Accounting and Processing Information), a working in real time computer system of central management of economy (Glushkov, 2004).

In order to acquire knowledge about the future state of the economy, we need to create (symbolic) algorithms that count similarly to the (real) algorithms according to which economic processes run, i.e. for the same past states, the predicted states are (almost) the same.

If we manage to create accurate algorithmic models of at least some economic processes, we will not necessarily be able to predict the results of real algorithms. There can be at least three reasons for this:

1. the symbolic algorithm poorly simulates the algorithm of economic processes,
2. the execution of the symbolic algorithm is slower than the real algorithm it is modeling,
3. the data transmission system on which the symbolic algorithm operates fails.

The world is already entwined with the web, the Internet, and although its development raises concerns about the possibility of privacy and, above all, freedom, especially in the face of manipulation, there is no sign of stopping it. Thanks to the global acquisition of up-to-date meteorological data, it becomes possible to better forecast

the weather. This is not the case with the economy. Is economic life more "capricious" than the weather, do we still not have access to sufficient data resources or do we not have symbolic algorithms that simulate the algorithms of economic life? However, we should also take into account the fact that the states of economy and the human behaviour are significantly interdependent and some predictions could be self-destructive. So far, the best people in the economy are those who have an intuition of management and have access to relevant data.

9.4 Machine motion and algorithmic evolution

Breakdown, "death", of a machine is a defect which may be caused by imperfect construction, faulty materials or workmanship. If man, the world of nature in general, were the work of a mechanical engineer, death would indicate a lack of engineering skills.

Naturalistically speaking, nature has created sophisticated structures such as organisms, living matter. The level of refinement is evidenced by the fact that man has still failed to create any form of living matter, and knowledge about life is—despite the tremendous achievements—still shallow. Every living creature is mortal, contrary to the expectations of these creatures. What has limited nature to produce individuals that are eternal? From a mechanistic perspective, we may ask what obstacles were to produce unbreakable machine/organism.

Death, the end of action, in the world of Galileo is not possible to describe without assuming some defect, some wear and tear, or exhaustion. The issue looks differently when viewed from the perspective of the algorithmic paradigm: an algorithm that has calculate the correct result stops. If the development of an organism is the implementation of an algorithm, then the dead of the organism indicates that

the algorithm completed the task for which it was written. From this perspective, death appears as fulfillment, as completion. In the world driven by algorithms, evolution is an algorithm. This evolutionary algorithm causes the death of imperfect organisms to make place for a more perfect ones. Those organisms that would achieve perfection could last forever.

Conceiving of organic life as an implementation of an algorithm is the leading idea of biocybernetics.

The immemorial problem of man is the question of free will. Zuse (Zenil, 2012, p.62–63) writes:

> I think the majority of researchers involved in the development of the computer have at some point in their lives, in one way or another considered the question of the relationship between human free will and causality.

Is there any satisfactory solution to this question, following the Turing paradigm?

How the good is the end of all our actions, as stated by Plato *Gorgias*: everything we do should be for the sake of what is good, and by Aristotle (1999, I.2):

> If, then, there is some end of the things we do, which we desire for its own sake (everything else being desired for the sake of this), and if we do not choose everything for the sake of something else (for at that rate the process would go on to infinity, so that our desire would be empty and vain), clearly this must be good and the chief good.

so also good would be the goal of algorithms.

In Newtonian physics, time and space are a boundless immutable "vessel" in which physical processes take place. In the case of the Tur-

ing paradigm, time and space are properties of algorithms. Relativity of time and space can be explained as determined by the execution of algorithms.

Following this line of thinking that evolution leads to improvement, will construction of a computer that is more perfect than man, a superintelligence, Čapek's robots, lead to a situation in which the algorithm of human life will terminate, because man has already fulfilled his task (Bostrom, 2014), or maybe?—as Kurzweil predicts (2005):

> The Singularity will allow us to transcend these limitations of our biological bodies and brains. We will gain power over our fates. Our mortality will be in our own hands. We will be able to live as long as we want (a subtly different statement from saying we will live forever). We will fully understand human thinking and will vastly extend and expand its reach. By the end of this century, the nonbiological portion of our intelligence will be trillions of trillions of times more powerful than unaided human intelligence.

9.5 The development of science

Science is a historical endeavor. Marciszewski describes it figuratively: Modern science is today an immeasurable ocean of knowledge, and the thought and output of Galileo, in conjunction with the pioneering work of Copernicus, is like the mouth of a river that waters gathered earlier for two millennia (Marciszewski and Stacewicz, 2011, p.232). If you ask where this current comes from, where and what are its sources, our river metaphor still holds true. It turns out that it is just like in nature. An identifiable spring is the beginning of this river, but it, in turn, has its origins in invisibly oozing streams buried in the grassland, without which our spring marked on the map would not exist.

We owe the achievements of science to our predecessors. Preached by John of Salisbury, echoing Bernard of Chartres, known for his attempts to reconcile the philosophy of Plato with that of Aristotle (Fairweather, 1956; John of Salisbury, 1159, III. CAP IV; 1955):

> *nos esse quasi nanos, gigantium humeris incidentes, ut possimus plura eis et remotiora videre, non utique proprii visus acumine, aut eminentia corporis, sed quia in altum subvehimur et extollimur magnitudine gigantea.*

Newton, whose *Philosophiæ Naturalis Principia Mathematica* (1687) opens the age of modern science, wrote to Robert Hooke (1675):

> If I have seen further, it is by standing on the shoulders of giants.

No generation has solved and—as the information philosophy justifies—will solve all problems, leaving them to future generations. This was already sensed by Newton (Westfall, 1983, p.643):

> To explain all nature is too difficult a task for any one man or even for any one age. Tis much better to do a little with certainty, & leave the rest for others that come after you, than to explain all things by conjecture without making sure of any thing.

Newton himself said (Brewster, 1855, p.407):

> I do not know what I appear to the world; but to myself I seem to have been only like a boy playing on a seashore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay undiscovered before me.

Creation of the science is similar to building medieval cathedrals. Everyone who participated in the construction of the cathedral had

different private goals and contributed differently to its construction, without being sure whether the cathedral would eventually be completed or what it would look like in the end. Nobody was sure when the construction would end.

Newton, whose bust at Trinity College has the inscription: *Qui genus humanum ingenio superavit* [there is no greater intellect among humans], who formulated (Newtonian) mechanics, which seemed to be the ultimate physical theory, did not believe that science could exhaust knowledge about the world. Today, thanks to information philosophy, we know that his gut feeling was right. The science of the digital age, as Marciszewski writes about, will be in a state of never-ending development, without exhausting all the consequences of the discovered truths (Marciszewski and Stacewicz, 2011).

Successive generations of researchers will expand, correct and explore knowledge resources, but there will still be areas that can be talked about—repeating after Emil du Bois-Reymond, a German physiologist, the belief *"ignoramus et ignorabimus"* [we do not know and know we will not], given in Leipzig at the lecture *Über die Grenzen des Naturerkennens* [On the limits of the knowledge of nature] at the Gesellschaft Deutscher Naturforscher und Ärtze (Du Bois-Reymond, 1872; 1882). He said that in the face of the puzzles of the material world, a nature researcher has long been used to the human reluctance to say his *'Ignoramus'* (we don't know). A look at past successes leads him to an unshakable awareness that, what he does not yet know, he could at least conditionally know, and one day he may. Faced with the mystery of what matter and force are and how they are conceivable, he must decide on a more difficult truth each time: *'Ignorabimus'* (we will not know):

> Gegen über den Rätseln der Körperwelt ist der Naturforscher längst gewöhnt, mit männlicher Entsagung sein 'Ignoramus'

> auszusprechen. Im Rückblick auf die durchlaufene siegreiche Bahn trägt ihn dabei das stille Bewußtsein, daß, wo er jetzt nicht weiß, er wenigstens unter Umständen wissen könnte, und dereinst vielleicht wissen wird. Gegenüber dem Rätsel aber, was Materie und Kraft seien, und wie sie zu denken vermögen, muß er ein für allemal zu dem viel schwerer abzugebenden Wahrspruch sich entschließen: 'Ignorabimus'.

David Hilbert (1900) did not agree with du Bois-Reymond's conviction, at least in mathematics. At a congress of mathematicians in Paris in 1900, he proclaimed that the inner voice says:

> Da ist das Problem, suche die Lösung. Du kannst sie durch reines Denken finden; denn in der Mathematik gibt es kein *Ignorabimus!*

At the end of his farewell speech in Königsberg on September 8, 1930, at a meeting of the Gesellschaft Deutscher Naturforscher und Ärtze, he claimed that (Hilbert, 1935, p.387; see also Smith, 2014):

> Wir müssen wissen,
> Wir werden wissen.

This belief was significant for the development of his research activities. The inscription of this content can be found on Hilbert's tombstone in the cemetery in Göttingen.

Hilbert's attempt to reject *Ignorabimus!* resulted in the creation of computer science and a justification—paradoxically—rejecting Hilbert's belief in the cognitive possibilities of formal methods.

10. Conclusions

Several comments and theses, even not completely developed and not satisfactorily justified, show that the Turing paradigm exceeds

the boundaries of formal sciences. It opens up new perspectives for research in the positive sciences. It also provides an opportunity for philosophical speculation about the world as made of algorithms. However, can we repeat after Konrad Zuse (Zenil, 2012, p.65) what he said in the middle of the 20th century?

> The concept of the computing universe is still just a hypothesis; nothing has been proved. However, I am confident that this idea can help unveil the secrets of nature.

Acknowledgments. I would like to express my gratitude to the anonymous reviewer for his valuable comments and suggestions.

Bibliography

Alexander de Villa Dei, 1839. Carmen de Algorismo. In: J.O. Halliwell-Phillipps, ed. *Rara Mathematica; or, a Collection of Treatises on the Mathematics and Subjects Connected with Them, from Ancient Inedited Manuscripts* [Online]. London: J. W. Parker; J.& J.J. Deighton; T. Stevenson, pp.73–83. Available at: <https://www.biodiversitylibrary.org/item/70672>.

Ares, J., Lara, J., Lizcano, D. and Martínez, M.A., 2018. Who Discovered the Binary System and Arithmetic? Did Leibniz Plagiarize Caramuel? *Science and Engineering Ethics* [Online], 24(1), pp.173–188. https://doi.org/10.1007/s11948-017-9890-6.

Aristotle, 1999. *Nicomachean Ethics* [Online] (W.D. Ross, Trans.). Kitchener, Ontario: Batoche Books. Available at: <https://socialsciences.mcmaster.ca/econ/ugcm/3ll3/aristotle/Ethics.pdf> [visited on 13 January 2023].

Babbage, C., 1864. *Passages from the Life of a Philosopher* [Online]. London: Longman, Green, Longman, Roberts, & Green. Available at: <http://djm.cc/library/Passages_Life_of_a_Philosopher_Babbage_edited.pdf> [visited on 13 January 2023].

Babbage, C., 2008. *Passages from the Life of a Philosopher, Ch. VIII* [Online]. Available at: <https://www.fourmilab.ch/babbage/lpae.html> [visited on 13 January 2023].

Bachelard, G., 2002. *The Formation of the Scientific Mind: A Contribution to a Psychoanalysis of Objective Knowledge* (M. McAllester Jones, Trans.), *Philosophy of science*. Manchester: Clinamen Press.

Bacon, R., 1912. *Part of the Opus Tertium of Roger Bacon including a fragment now printed for the first time* [Online]. Ed. by A. Little. Vol. 4. Aberdeen: Aberdeen University Press. Available at: <http://capricorn.bc.edu/siepm/DOCUMENTS/BACON/Bacon_partofopustertium(ed.Little).pdf> [visited on 13 January 2023].

Bacon, R., 2010. Mathematical Science. In: J.H. Bridges, ed. *The Opus Majus of Roger Bacon* [Online]. Cambridge University Press, pp.97–404. https://doi.org/10.1017/CBO9780511709661.006.

Bocheński, J.M., 1994. *Sto zabobonów: krótki filozoficzny słownik zabobonów*. Kraków: Philed.

Bostrom, N., 2014. *Superintelligence: Paths, Dangers, Strategies*. 1st ed. Oxford: Oxford University Press.

Brewster, D., 1855. *Memoirs of the Life, Writings, and Discoveries of Sir Isaac Newton* [Online]. Edinburgh: T. Constable and Co. Available at: <http://archive.org/details/memoirslifewrit02brewgoog> [visited on 13 January 2023].

Burks, A., Goldstine, H. and von Neumann, J., 1946. *Preliminary Discussion of the Logical Design of an Electronic Computing Instrument* [Online]. U.S. Army Ordnance Department & Institute for Advanced Study. Available at: <https://www.ias.edu/sites/default/files/library/Prelim_Disc_Logical_Design.pdf> [visited on 17 January 2023].

Burks, A., Goldstine, H. and von Neumann, J., 1987. Preliminary discussion on the logical design of an electronic computing instrument. In: W. Aspray and A.W. Burks, eds. *Papers of John von Neumann on Computing and Computer Theory* [Online]. Vol. 12. Cambridge, MA: MIT Press, pp.97–142. Available at: <http://archive.org/details/papersofjohnvonn00vonn> [visited on 13 January 2023].

Cantor, M., 1865. Über einen Codex des Klosters Salem. *Zeitschrift für Mathematik und Physik* [Online], 10, pp.1–16. https://doi.org/10.11588/heidok.00012869.

Caramuel y Lobkowitz, J., 1670. *Ioannis Caramuelis mathesis biceps: vetus et nova*. Campaniae: in Officina Episcopali.

Chaitin, G.J., 2004a. *Leibniz, Randomness and the Halting Probability* [Online]. arXiv. https://doi.org/10.48550/arXiv.math/0406055.

Chaitin, G.J., 2004b. Meta Math! The Quest for Omega. *arXiv:math/0404335* [Online], p.150. Available at: <http://arxiv.org/abs/math/0404335> [visited on 11 February 2015].

Chaitin, G., 2007. Epistemology as Information Theory: From Leibniz to Ω. In: G. Dodig Crnkovic, ed. *Computation, Information, Cognition – The Nexus and The Liminal*. Newcastle, UK: Cambridge Scholars Pub., pp.2–17.

Cherry, S., 2017. *The Reason of Reason: How Reason, Logic, and Intelligibility Together are Evidence for God*. Canterbury: Telos Publishing.

Condorcet, J.-A.-N.d.C., 1785. *Essai sur l'application de l'analyse à la probabilité des décisions rendus à la pluralité des voix* [Online]. Paris: Imprimerie Royale. Available at: <http://archive.org/details/essaisurlapplica00cond> [visited on 13 January 2023].

Copeland, J., Shagrir, O. and Sprevak, M., 2016. Is the whole universe a computer? In: J. Copeland, J. Bowen, M. Sprevak and R. Wilson, eds. *The Turing Guide* [Online]. Oxford; New York: Oxford University Press, pp.445–462. https://doi.org/10.1093/oso/9780198747826.003.0054.

Couturat, L., 1901. *La logique de Leibniz d'après des documents inédits* [Online]. Paris: F. Alcan. Available at: <http://archive.org/details/lalogiquedeleib00coutgoog> [visited on 13 January 2023].

Davis, M., 2001. *Engines of Logic: Mathematicians and the Origin of the Computer*. New York: Norton.

De Morgan, A., 1872. *A Budget of Paradoxes* [Online]. London: Longmans, Green, and Co. Available at: <http://archive.org/details/budgetofparadoxe00demorich> [visited on 13 January 2023].

Dodig-Crnkovic, G., 2013. Alan Turing's Legacy: Info-computational Philosophy of Nature. In: G. Dodig-Crnkovic and R. Giovagnoli, eds. *Com-*

puting Nature: Turing Centenary Perspective [Online], *Studies in Applied Philosophy, Epistemology and Rational Ethics*. Berlin; Heidelberg: Springer, pp.115–123. https://doi.org/10.1007/978-3-642-37225-4_6.

Dodig-Crnkovic, G., 2022. In search of a common, information-processing, agency-based framework for anthropogenic, biogenic, and abiotic cognition and intelligence. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (73), pp.17–46.

Donahue, W.H., 1993. Kepler's First Thoughts on Oval Orbits: Text, Translation, and Commentary. *Journal for the History of Astronomy* [Online], 24(1-2), pp.71–100. https://doi.org/10.1177/002182869302400103.

Du Bois-Reymond, E.H., 1872. *Über die Grenzen des Naturerkennens: Ein Vortrag in der zweiten öffentlichen Sitzung der 45. Versammlung deutscher Naturforscher und ärzte zu Leipzig am 14. August 1872*. 1st ed. Leipzig: Veit & Co.

Du Bois-Reymond, E.H., 1882. *Über die grenzen des Naturerkennens: die sieben Welträthsel zwei Vorträge* [Online]. Leipzig: Veit. Available at: <https://wellcomecollection.org/works/srwgn7yp> [visited on 13 January 2023].

Duch, W., 2012. Neuronauki i natura ludzka. In: M. Słomka, ed. *Nauki przyrodnicze a nowy ateizm*, *Filozofia Przyrody i Nauk Przyrodniczych*, 8. Lublin: Wydawnictwo KUL, pp.79–122.

Duch, W., 2015. *Katolicyzm po kognitywistyce: o nową teologię umysłu*. Available at: <http://teologia.deon.pl/katolicki-obraz-natury-ludzkiej-i-nauki-kognitywne/> [visited on 17 January 2023].

Duch, W., 2017. Why minds cannot be received, but are created by brains. *Scientia et Fides* [Online], 5(2), pp.171–198. Available at: <https://apcz.umk.pl/SetF/article/view/SetF.2017.014> [visited on 17 January 2023].

Eddington, A.S., 2014. *The Nature of the Physical World: Gifford Lectures of 1927: An Annotated Edition*. Ed. by H.G. Callaway. Cambridge: Cambridge Scholars Publishing.

Fairweather, E.R., 1956. *A scholastic miscellany: Anselm to Ockham*. Philadelphia: Westminster Press.

Floridi, L., 2002. What is the Philosophy of Information? *Metaphilosophy* [Online], 33(1-2), pp.123–145. https://doi.org/10.1111/1467-9973.00221.

Floridi, L., 2008. A defence of informational structural realism. *Synthese* [Online], 161(2), pp.219–253. https://doi.org/10.1007/s11229-007-9163-z.

Floridi, L., 2009. The Information Society and Its Philosophy: Introduction to the Special Issue on "The Philosophy of Information, Its Nature, and Future Developments". *The Information Society* [Online], 25(3), pp.153–158. https://doi.org/10.1080/01972240902848583.

Galileo Galilei, 1615. *Letter to Madame Christina of Lorraine, Grand Duchess of Tuscany* (S. Drake, Trans.). Available at: <https://inters.org/galilei-madame-christina-Lorraine> [visited on 17 January 2023].

Galileo Galilei, 1623. *The Assayer (Il Saggiatore)* [Online] (S. Drake, Trans.). Available at: <https://web.stanford.edu/~jsabol/certainty/readings/Galileo-Assayer.pdf> [visited on 13 January 2023].

Glushkov, V., 2004. *Chto skazhet istorija? (in russian)*. Available at: <https://web.archive.org/web/20100426165336/http://www.situation.ru/app/j_art_333.htm> [visited on 17 January 2023].

Guthrie, K.S. and Fideler, D., eds., 1987. *The Pythagorean Sourcebook and Library: An Anthology of Ancient Writings Which Relate to Pythagoras and Pythagorean Philosophy* [Online]. Grand Rapids, MI: Phanes Press. Available at: <https://ia801704.us.archive.org/17/items/guthrie-1987-the-pythagorean-sourcebook-and-library/Guthrie%201987%20The%20Pythagorean%20Sourcebook%20and%20Library.pdf> [visited on 13 January 2023].

Hall, A.R., 1956. *The Scientific Revolution, 1500-1800: The Formation of the Modern Scientific Attitude* [Online]. Boston: Beacon Press. Available at: <http://archive.org/details/scientificrevolu00hall> [visited on 13 January 2023].

Heller, M., 2013. *Bóg i nauka: moje dwie drogi do jednego celu* (E. Nicewicz-Staszowska, Trans.). 1st ed. Kraków: Copernicus Center Press.

Heller, M., 2014. *Granice nauki*. Kraków: Copernicus Center Press.

Heschmeyer, J., 2012. *Two Interesting Arguments for God: Intelligibility & Desire*. Available at: <http://shamelesspopery.com/two-interesting-arguments-for-god-intelligibility-desire/> [visited on 17 January 2023].

Hilbert, D., 1900. Mathematische Probleme. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* [Online]. Göttingen: Vandenhoeck & Ruprecht, pp.253–297. Available at: <https://de.wikisource.org/wiki/Mathematische_Probleme> [visited on 13 January 2023].

Hilbert, D., 1935. Naturerkennen und Logik. *Gesammelte Abhandlungen (Dritter Band)* [Online]. Berlin: Verlag von Julius Springer, pp.378–087. Available at: <https://gdz.sub.uni-goettingen.de/id/PPN237834022> [visited on 13 January 2023].

Hobbes, T., 1651. *Leviathan, Or, The Matter, Form, and Power of a Commonwealth, Ecclesiastical and Civil*. London: Andrew Crooke.

Hodges, A., 1997. *Turing: A Natural Philosopher*. 1st ed., *The great philosophers*, 3. London: Phoenix.

Ineichen, R., 2008. Leibniz, Caramul, Harriot und das Dualsystem. *Mitteilungen der Deutschen Mathematiker-Vereinigung* [Online], 16(1), pp.12–15. https://doi.org/10.1515/dmvm-2008-0009.

Infeld, L., 1980. *Quest: An Autobiography*. 2d ed. New York, NY; London: Chelsea Pub. Co.

Isidore of Seville, 1911. *Etymologiarvm sive Originvm libri XX;* [Online]. Ed. by W.M.(.M. Lindsay. Oxonii: e typographeo Clarendoniano. Available at: <http://archive.org/details/isidorihispalen00lindgoog> [visited on 13 January 2023].

John of Salisbury, 1159. *Metalogicus* [Online]. Available at: <http://www.logicmuseum.com/wiki/Authors/John_of_Salisbury/Metalogicon> [visited on 14 January 2023].

John of Salisbury, 1955. *The Metalogicon of John of Salisbury* (D.D. McGarry, Trans.). Berkeley; Los Angeles: University of California Press.

Knuth, D.E., 1997. *The Art of Computer Programming: Vol 1. Fundamental Algorithms*. 3rd ed. Reading, MA: Addison-Wesley.

Kopania, J., 2018. Leibniz i jego Bóg. Rozważania z Voltaire'em w tle. *Studia z Historii Filozofii* [Online], 9(3), p.69. https://doi.org/10.12775/szhf.2018.031.

Krajewski, S., 2012. The ultimate strengthening of the Turing Test? *Semiotica* [Online], 2012(188), pp.203–218. https://doi.org/10.1515/sem-2012-0014.

Krajewski, S., 2020. On the Anti-Mechanist Arguments Based on Gödel's Theorem. *Studia Semiotyczne* [Online], 34(1), pp.9–56. https://doi.org/10.26333/sts.xxxiv1.02.

Kuhn, T.S., 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kuhn, T.S., 1974. Second thoughts on paradigms. In: F. Suppe, ed. *The Structure of Scientific Theories*. Urbana: University of Illinois Press, pp.459–482.

Kurzweil, R., 2005. *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking.

Leibniz, G.W., 1666. *Dissertatio de arte combinatoria, in qua ex arithmeticae fundamentis complicationum ac transpositionum doctrina nouis praeceptis exstruitur ... noua etiam Artis meditandis, seu Logicae inuentionis semina sparguntur. Praefixa est synopsis totius tractatus, & additamenti loco demonstratio existentiae Dei, ad mathematicam certitudinem exacta autore Gottfredo Guilielmo Leibnüzio Lipsensi...* [Online]. Lipsiae [Leipzig]: Joh. Simon. Fickium et Joh. Polycarp. Senboldum. Available at: <http://archive.org/details/ita-bnc-mag-00000844-001> [visited on 14 January 2023].

Leibniz, G.W., 1697. *Brief an den Herzog von Braunschweig-Wolfenbüttel Rudolph August, 2. Januar 1697* [Online]. Available at: <http://www.fh-augsburg.de/~harsch/germanica/Chronologie/17Jh/Leibniz/lei_bina.html> [visited on 14 January 2023].

Leibniz, G.W., 1890a. Dialogus. In: C.I. Gerhardt, ed. *Die philosophischen Schriften von Gottfried Wilhelm Leibniz, herausg. von C.J. Gerhardt. (Vol. 7)* [Online]. Berlin: Weidmann, pp.190–193. Available at: <http://purl.ox.ac.uk/uuid/f9e9db855bb7460498e270f024a7d8dc> [visited on 14 January 2023].

Leibniz, G.W., 1890b. *Die philosophischen Schriften von Gottfried Wilhelm Leibniz, herausg. von C.J. Gerhardt. (Vol. 7)* [Online]. Ed. by C.I. Gerhardt. Berlin: Weidmann. Available at: <http://purl.ox.ac.uk/uuid/f9e9db855bb7460498e270f024a7d8dc> [visited on 14 January 2023].

Leibniz, G.W., 1929. On calculating machine [Machina arithmetica in qua non additio tantum et subtractio sed et multiplicatio nullo, divisio vero paene nullo animi labore peragantur]. In: D.E. Smith, ed. *A Source Book in Mathematics* [Online] (M. Kormes, Trans.). New York: McGraw-Hill Book Co., pp.173–181. Available at: <http://archive.org/details/sourcebookinmath00smit> [visited on 14 January 2023].

Leibniz, G.W., 1979. De progressione dyadica (1679) [with German translation]. *Herrn von Leibniz' Rechnung mit Null und Eins* [Online]. Berlin; München: Siemens Aktiengesellshaft, pp.46–47. Available at: <http://www.heenes.de/ro/material/leibniz/leibniz.pdf> [visited on 13 January 2023].

Leibniz, G.W., 1990. *Leibniz korrespondiert mit China: der Briefwechsel mit den Jesuitenmissionaren (1689-1714)*. Ed. by R. Widmaier. Frankfurt am Main: V. Klostermann.

Lesne, A., 2007. The discrete versus continuous controversy in physics. *Mathematical Structures in Computer Science* [Online], 17(2), pp.185–223. https://doi.org/10.1017/S0960129507005944.

Ligonnière, R., 1992. *Prehistoria i historia komputerów: od początków rachowania do pierwszych kalkulatorów elektronicznych* (R. Dulinicz, Trans.). Wrocław: Zakład Narodowy im. Ossolińskich.

Marciszewski, W., ed., 1981. *Dictionary of Logic as Applied in the Study of Language: Concepts—Methods—Theories*, *Nijhoff International Philosophy Series*, vol. 9. Hague; Boston; London: Martinus Nijhoff.

Marciszewski, W. and Stacewicz, P., 2011. *Umysł-komputer-świat. O zagadce umysłu z informatycznego punktu widzenia* [Online]. Akademicka Oficyna Wydawnicza EXIT. Available at: <http://libra.ibuk.pl/book/101353> [visited on 6 March 2014].

Martin, M. and Augustine, K., eds., 2015. *The Myth of an Afterlife: The Case Against Life After Death*. Lanham, MD: Rowman & Littlefield.

Mazur, J., 2014. *Enlightening Symbols: A Short History of Mathematical Notation and Its Hidden Powers*. Princeton: Princeton University Press.

Menninger, K., 1958. *Zahlwort und Ziffer: eine Kulturgeschichte der Zahl*. 2., neubearb. und erw. Aufl. Göttingen: Vandenhoeck & Ruprecht.

Menninger, K., 1969. *Number Words and Number Symbols; a Cultural History of Numbers* [Online] (P. Broneer, Trans.). Cambridge, MA: MIT Press. Available at: <http : / / archive . org / details / numberwordsnumbe00menn> [visited on 14 January 2023].

von Neumann, J., 1958. *The Computer and the Brain*. 1st ed., *Mrs. Hepsa Ely Silliman Memorial Lectures*. New Haven: Yale University Press.

von Neumann, J., 1963. The general and logical theory of automata. In: A.H. Taub, ed. *John Von Neumann Collected Works: Volume V: Design of Computers, Theory of Automata and Numerical Analysis* [Online]. Pergamon Press, pp.288–328. [visited on 16 January 2023].

von Neumann, J., 2012. *The Computer & the Brain*. 3rd ed. New Haven, CT; London: Yale University Press.

von Neumann, J. and Burks, A.W., 1966. *Theory of Self-Reproducing Automata*. Urbana, IL; London: University of Illinois Press.

Newton, I., 1675. *Isaac Newton letter to Robert Hooke [ID 9792]*. Available at: <https://digitallibrary.hsp.org/index.php/Detail/objects/9792> [visited on 14 January 2023].

Newton, I., 1687. *Philosophiae Naturalis Principia Mathematica* [Online]. Londini; [London]: J. Societatis Regiae ac Typis J. Streater. Available at: <https://books.google.pl/books?id=uuRHAQAAMAAJ> [visited on 14 January 2023].

Shannon, C.E., 1958. Von Neumann's contributions to automata theory. *Bulletin of the American Mathematical Society* [Online], 64(3), pp.123–129. https://doi.org/10.1090/S0002-9904-1958-10214-1.

Shirley, J.W., 1951. Binary Numeration before Leibniz. *American Journal of Physics* [Online], 19, pp.452–454. https://doi.org/10.1119/1.1933042.

Sibley, A., 2013. Lessons from Augustine's De Genesi ad Litteram—Libri Duodecim. *Journal of Creation*, 27(2), pp.71–77.

Smith, J., 2014. David Hilbert's Radio Address. *Convergence* [Online]. https://doi.org/10.4169/convergence20140202.

Sonar, T., 2018. *The History of the Priority Dispute between Newton and Leibniz* [Online]. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-72563-5.

Strogatz, S.H., 2019. *Infinite Powers: How Calculus Reveals the Secrets of the Universe*. Boston: Houghton Mifflin Harcourt.

Swade, D., 2002. *The Difference Engine: Charles Babbage and the Quest to Build the First Computer* [Online]. New York: Penguin Books. Available at: <http://archive.org/details/differenceengine00doro> [visited on 14 January 2023].

Swetz, F.J., 2003. Leibniz, the Yijing, and the Religious Conversion of the Chinese. *Mathematics Magazine* [Online], 76(4), pp.276–291. https://doi.org/10.2307/3219083.

Swift, J., 1892. *Gulliver's Travels into Several Remote Nations of the World* [Online]. Ed. by D. Price. London: George Bell amd Sons. Available at: <http://%20www.gutenberg.org/ebooks/829> [visited on 14 January 2023].

Tegmark, M., 2008. The Mathematical Universe. *Foundations of Physics* [Online], 38(2), pp.101–150. https://doi.org/10.1007/s10701-007-9186-9.

Tegmark, M., 2014. *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. 1st ed. New York: Alfred A. Knopf.

Trzęsicki, K., 2006a. From the Idea of Decidability to the Number 'Omega'. *Studies in Logic, Grammar and Rhetoric* [Online], 9(22), pp.73–142. Available at: <http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.81a2f1db-908d-3270-a7fd-f231bdf84be6> [visited on 14 January 2023].

Trzęsicki, K., 2006b. Leibniza idea systemu binarnego. In: J. Kopania and H. Święczkowska, eds. *Filozofia i myśl społeczna XVII wieku*. Białystok: Wydawnictwo Uniwersytetu, pp.183–203.

Trzęsicki, K., 2006c. Leibnizjańskie inspiracje informatyki. *Filozofia Nauki* [Online], 14(3 (55)), pp.21–48. Available at: <https://www.ceeol.com/search/article-detail?id=97255> [visited on 14 January 2023].

Trzęsicki, K., 2016. Can AI Be Intelligent? *Studies in Logic, Grammar and Rhetoric* [Online], 48(1), pp.103–131. https://doi.org/10.1515/slgr-2016-0058.

Trzęsicki, K., 2020a. Idea of Artificial Intelligence. *Studia Humana* [Online], 9(3/4), pp.37–65. https://doi.org/10.2478/sh-2020-0027.

Trzęsicki, K., 2020b. Idea sztucznej inteligencji. *Filozofia i Nauka. Studia Filozoficzne i Interdyscyplinarne* [Online], 8(1), pp.69–96. https://doi.org/10.37240/FiN.2010.8.1.3.

Turing, A.M., 1937. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* [Online], 42 (series 2)(1), pp.230–265. https://doi.org/10.1112/plms/s2-42.1.230.

Turing, A.M., 1950. Computing Machinery and Intelligence. *Mind* [Online], 59(236), pp.433–460. https://doi.org/10.1093/mind/LIX.236.433.

Turing, A.M., 1952. The Chemical Basis of Morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* [Online], 237(641), pp.37–72. Available at: <http://www.jstor.org/stable/92463> [visited on 4 November 2015].

Westfall, R.S., 1983. *Never at Rest: A Biography of Isaac Newton*. Cambridge: Cambridge University Press.

Wheeler, J.A., 1989. Information, Physics, Quantum: The Search for Links. *Proceedings of the 3rd International Symposium Foundations of Quantum Mechanics in the Light of New Technology: Central Research Laboratory, Hitachi, Ltd., Kokubunji, Tokyo, Japan, August 28-31, 1989*. Tokyo: Physical Society of Japan, pp.354–368.

Whitehead, A.N., 1911. *An Introduction to Mathematics* [Online]. London; New York: Williams & Northgate; H. Holt. Available at: <http://archive.org/details/introductiontoma00whitiala> [visited on 16 January 2023].

Wolfram, S., 2002. *A New Kind of Science* [Online]. Champaign, IL: Wolfram Media. Available at: <http://www.wolframscience.com/nks/> [visited on 16 January 2023].

Wolfson, H.A., 1962. The Problem of the Souls of the Spheres from the Byzantine Commentaries on Aristotle Through the Arabs and St. Thomas to Kepler. *Dumbarton Oaks Papers* [Online], 16, pp.65–93. https://doi.org/10.2307/1291158.

Wouk, H., 2010. *The Language God Talks: On Science and Religion*. 1st ed. New York, NY: Little, Brown and Co.

Zenil, H., 2012. Afterword to Konrad Zuse's Calculating Space. In: A. German and H. Zenil, eds. *A Computable Universe: Understanding Computation & Exploring Nature As Computation (with a Foreword by Sir Roger Penrose)* [Online]. Singapore; Hackensack, NJ: World Scientific, pp.787–794. Available at: <https://www.mathrix.org/zenil/ZuseCalculatingSpace-GermanZenil.pdf> [visited on 13 January 2023].

Zuse, K., 1967. Rechnender Raum. *Elektronische Datenverarbeitung* [Online], 8, pp.336–344. Available at: <https://www.informationphilosopher.com/solutions/%20scientists/zuse/Rechnender_Raum.pdf> [visited on 16 January 2023].

Zuse, K., 1969. *Rechnender Raum*. Vol. 1, *Schriften zur Datenverarbeitung*. Braunschweig: Vieweg.

Zuse, K., 2010. *Der Computer – Mein Lebenswerk* [Online]. 5., unveränd. Aufl. Berlin; Heidelberg: Springer. https://doi.org/10.1007/978-3-642-12096-1.

Zuse, K., 2012a. Calculating Space (Rechnender Raum). *A Computable Universe* [Online]. WORLD SCIENTIFIC, pp.729–786. https://doi.org/10.1142/9789814374309_0036.

Zuse, K., 2012b. *Calculating space (Rechnender Raum)* [Online]. Ed. by A. German and H. Zenil. Available at: <https://www.mathrix.org/zenil/ZuseCalculatingSpace-GermanZenil.pdf> [visited on 16 January 2023].

Review articles

Artykuły recenzyjne

Beyond epistemic concepts of information: The case of ontological information as philosophy in science

Paweł Polak
Pontifical University of John Paul II in Kraków, Poland

Roman Krzanowski, *Ontological information: information in the physical world, series: World Scientific series in information studies*, 13, Hackensack, NJ: World Scientific 2022, pp. xii+264.

The concept of information plays and important role in science and philosophy, as well as in everyday life, such that it is now hard to imagine that this concept has been only adopted in the late 1940s. Many scientists find it even harder to believe that the concept of information can involve anything other than communication processes, and this is almost certainly due to Claude Shannon's Theory of Communication (TOC) (1949) that entered the canon of unquestionable modern scientific knowledge. Unquestionably accepting Shannon's concept of a measure of information entropy as the definition of information encourages a scholar to treat TOC as scientific dogma.[1] However,

[1] Shannon himself tried to warn against abusing his theory of communication (Shannon, 1956), though apparently unsuccessfully. He called against using the theory as a source of hypotheses in other scientific disciplines. However, the opposite has happened—information metaphors have become unquestionable theoretical core of many modern concepts. This fact should not come as a surprise, because if few people read the original work contenting themselves only with its processed results, it is difficult to suppose that anyone outside a handful of specialists in the history of computing read Shannon's critical remarks very rarely cited in the literature.

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 335–345

studying the original work of Shannon and Weaver (1964, p.3), we realize that Shannon was primarily interested in communication engineering of digital signals (signal recovery, noise, optimal coding of signal), and the concepts of information and entropy have been borrowed by him from works of physicists like Ludwig Boltzmann, Leo Szilard, and John von Neumann.

After almost 80 years of being around, information is an elusive concept with manifold meanings; Krzanowski (2022) referred to more than 300 definitions of information, As information plays such an important role in contemporary societies, technology and science, it seems only logical that the efforts to clarify the meaning of information should never be abandoned. And this is precisely what Krzanowski's book is about.

However, while most of the published works on information see information through Shannon's lenses the focus of Roman Krzanowski's book (Krzanowski, 2022) is physical information i.e., information that is not associated with knowledge or communication (Shannon's legacy), and that it is a part of nature as other physical phenomena are; the conceptualization of information has not been widely accepted by the scientific community.

The first significant step was to distinguish between concepts of epistemic information (as found in Shannon's theory) from theories about ontological information. The author's second step was rather than constructing a concept of information *a priori*, to proceed in the spirit of the Kraków school of philosophy in science (Heller, 2019; Polak, 2019; Polak and Trombik, 2022). Krzanowski searches for the meaning of ontological information g attributed to it by reserachers and tries to understand the philosophical basis for using such concept. This is no coincidence, because in the pages of the associated journal "Philosophical Problems in Science/Zagadnienia Filozoficzne

w Nauce" (ZFN), the problem of information in science has been discussed from the very beginning to the current day (Turek, 1978; 1981; Krzanowski, 2017; 2020).

The book is divided over seven chapters. The first chapter introduces the several definitions of information, often contradictory, demonstrating difficulty in accurately capturing the essence of this concept. Throughout the chapter the author gradually builds the conceptual base for his thesis and carefully justifies all his decisions. Of course, it is possible to disagree with Krzanowski on many issues, but one must admit that he tries to be very consistent, meaning that the deliberations as a whole constitute a valuable analysis of the concept of information. Even if one disagrees with the author's detailed claims, one would still concede that this book takes on an intriguing intellectual challenge and makes a significant contribution to organizing and illuminating the discussion around the concept of information.

At a time when authors mainly value their own originality, the work of Krzanowski has the characteristics of the best classical philosophy, which built its solutions on critical struggles with the heritage of tradition. It undoubtedly contributes to modern analytic philosophy, but the author's approach is to not simply copy contemporary models but instead creatively draw from various traditions, including Polish analytic philosophy. Although the author is far removed from the theses of Aristotelianism and scholasticism, his perfectly organized, methodical criticism and consistency and his precise argumentation is reminiscent of the style of Thomas Aquinas. More importantly, though, Krzanowski is not pragmatophobic, instead boldly pursuing solutions and seeking his own synthesis in the thicket of proposals. This method certainly sets this book apart from most works on the concept of information.

The crucial concept of ontological information is characterized as "a physical phenomenon" (Krzanowski, 2022, p.6). The author assumes that "this information is perceived as a structure, organization, or form of natural and artificial (artifacts) objects." He also defines this information as being objective and mind-independent while simultaneously clarifying all the concepts involved and trying to provide an argument for every claim. He also warns that ontological information is a metaphysical concept and contributes to contemporary analytical metaphysics.

Krzanowski's analyses start with some intuitions about ontological information. He reconstructs ideas from dispersed quotations, much like how historians of philosophy deal with pre-Socratic philosophy. The methodology is similar because the concept of ontological information frequently manifests itself in the form of dispersed brief remarks.

The collection of these brief remarks by scientists is combined with a careful interpretation and an attempt to reconstruct the philosophical intuitions they contain, the tasks that are the subject of the second chapter. This intriguing and inspiring journey passes through a variety of ideas, culminating in the formulation of the eight main intuitions about ontological information that run through the scientific literature.

The third chapter analyzes the existing philosophical conceptions of ontological information, even though they do not usually refer to it using this term. We can find concepts coined by representatives of different disciplines from various countries, such as Carl von Weizsäcker, Krzysztof Turek, Stefan Mynarski, John Collier, Tom Stonier, Michał (Michael) Heller, Gordana Dodig Crnkovic, César Hildago, Thomas

Nagel, Jacek Jadacki, and Anna Brożek. Krzanowski summarizes these concepts into 11 claims that are explained in detail (Krzanowski, 2022, pp.86–93).

The aforementioned intuitions and claims serve as a foundation for synthesizing the concept of ontological information in the fourth chapter, with Krzanowski ultimately reducing it to three claims:

- (EN) Information has no meaning, but meaning is derived from information by a cognitive agent.
- (PE) Information is a physical phenomenon.
- (FN) Information is responsible for the organization of the physical world.

This is then followed by two corollaries:

- (C1) Information is quantifiable.
- (C2) Changes in the organization of physical objects can be denoted as a form of computation or information processing.

After the critical discussion, Krzanowski posits that these three properties and two corollaries are indispensable for the definition and understanding of the concept of ontological information. This set of properties has a hypothetical status, and this conceptualization is relative to actual science, so it is open to future changes together with the entirety of scientific knowledge.

The fifth chapter is devoted to broadly analyzing the problem of ontological and epistemological aspects of the concept of information. Krzanowski needs to adopt this perspective to further clarify the concept of ontological information. The analysis shows that while both concepts of information are required to account for the full

spectrum of interpretations for information, ontological information appears to be more fundamental because it can serve as a carrier of epistemic information.

In the following chapter, Krzanowski moves onto applications and interpretations of ontological information. He critically discusses the concept of an "infon" and data as basic concepts for defining information. The author claims that his conceptualization is more fundamental and better explains the source of epistemic information. Furthermore, Krzanowski attempts to resolve the dilemma of the contradictory abstract and concrete natures for information, and this is another original and inspiring aspect of his book. The example application of ontological information is Krzanowski's original concept of Minimal Information Structural Realism (first introduced in Krzanowski, 2017). Also of interest is the consideration about possibly applying his conceptualization of ontological information to Popper's Three Worlds and Mark Burgin's General Theory of Information. Finally, Krzanowski applies Perzanowski's ontology to build an ontological foundation for the concept of ontological information.

The final chapter gathers together all the observations from the book, summarizes the key findings and conclusions, and brings up some selected criticisms of ontological information. (Krzanowski probably intentionally avoids the most common dogmatic critiques, regarding them as not being suitable for philosophical consideration.) Finally, through nine questions, the author reveals some perspectives for future research into ontological information. Each question opens up a new field that could be a subject of a new study.

It is worth mentioning that the book has been carefully prepared from an editorial perspective. It has not, however, been spared from minor inaccuracies, such as the fact that Krzysztof Turek received his doctorate from the Pontifical Academy of Theology in Kraków,

which only transformed into the Pontifical University of John Paul II in Kraków many years later. In addition, qualifying Jacek Jadacki as a computer scientist (Krzanowski, 2022, p.45) rather than a philosopher and pianist is not only untrue—it is also inconsistent with the rest of the work. Nevertheless, these minor glitches do not significantly impair this important consideration of ontological information. Unfortunately, many more, albeit minor, errors can be found in the bibliography, especially in the Polish titles of works. This may hinder any search for the cited works in databases. Moreover, in some cases, works that have long been out of print, even five years ago, are still marked as being in print.

Assessing philosophical import of the book we may begin by noting that the book is relatively new, but published works have already used the ideas within it. Work that is worthy of mentioning here is that of the philosophy of information specialist Mark Burgin (Burgin and Mikkilineni, 2022). The ideas are also reflected in this issue of *Philosophical Problems in Science* / ZFN (Mścisławski, 2022). We should also emphasize here that the ideas presented in the reviewed book have resulted from the longer, critical reflection conducted by Krzanowski. This is especially true of the concepts of physical information and information, which were originally treated as being synonymous but have now been distinguished in the book, and this division has been well justified.

After reading the book, many questions can be raised, but to be fair to the author, they should be posed with great precision and care. There are certainly questions about whether the book finally resolves the problem of defining information or whether it finally explains the nature of information, but these would be misplaced questions. Indeed, it would be unacceptable for science to achieve these goals through *a priori* considerations, so if we adopt the scientific perspective, we

must accept that we cannot provide definitive answers. Nevertheless, this does not mean that we must remain mute on the subject. On the contrary, we can say much about how the concept of information functions in modern science. Of course, Krzanowski's book only addresses the issue of ontological information, because philosophers have paid far too little attention to it. Indeed, the fame of Shannon's work on the theory of communication (often interpreted as the theory of information) has all too often led to an atrophy of criticism and a limited vision for the nature of information. For this reason, the reviewed book makes a valuable contribution to the discussion, and it not only reconstructs the concept of ontological information that is actually used in science but also critically evaluates it.

The book formulates a set of properties of ontological information. This is the first attempt of its kind, and most importantly, it does not start from the author's arbitrary ideas but rather tries to deal with the thicket of intuitions and conceptualizations put forward in scholarly publications. The task is hard as scientists often hide their ignorance behind imprecise statements. After all, they are not professional philosophers, and in this task, which is secondary to their research, they may easily fall fowl of various errors or inaccuracies. This is perhaps why Einstein pointed out that one should pay attention to what scientists actually do rather than what they say about it. If this is indeed the case, then I would like to propose an important area to develop research into the concept of ontological information, namely to investigate how it is actually used in scientific research.

Thus, it becomes necessary to go beyond the mere declarations and conceptualizations made by scholars and look at how the concept is used in explanatory structures and other aspects of research practice, at least if such contexts can be identified. Nevertheless, it should be

noted that at the time of writing, such a task was essentially impossible given the scarcity of studies using the concept of ontological information.

Of course, the author's assertion that the concept of ontological information is a metaphysical concept should be borne in mind. Thus, the proposed line of research is also a proposed case study of how metaphysics interacts with the sciences using the example of the modern concept of information. Such studies are also lacking, yet they could be of value to all philosophers of science who do not share the extreme anti-metaphysical position.

We can hope that this reviewed book and the study areas suggested for continuing research should have some impact on science. By stripping certain ideological trappings from the concept of information, broadening the perspectives (escaping Shannon's shadow) and making some necessary clarifications, scientists should certainly be able to develop new lines of research more effectively. Besides, the first harbingers of change have already appeared, such as the work flowing from the Kraków scientific community (Bielecki and Schmittel, 2022), which is based on the work of Krzanowski. Let us hope that other works using such a "purified" concept of ontological information in scientific practice will soon appear, because it will open up a new yet important step in the philosophy of information, namely research into ontological information in scientific explanatory structures.

Acknowledgments The author thanks Łukasz Mścisławski for the in-depth discussions of the concept of information. His article (Mścisławski, 2022) published in this issue of ZFN is the first critical analysis of the concepts presented by R. Krzanowski; Łukasz Mścisławski's ideas significantly contributed to a this review.

Abstract

This review article discusses the book of Roman Krzanowski, *Ontological Information: Information in the Physical World*, which is published by World Scientific. Krzanowski's book makes a very important contribution to the contemporary discussion about the nature of information. The author analyzes the concept of ontological information and its uses in the works of scientists from various disciplines, resulting in an innovative and inspiring analysis that every philosopher involved in the philosophy of information should read.

Keywords

ontological information, philosophy in science, Mark Burgin

Bibliography

Bielecki, A. and Schmittel, M., 2022. The Information Encoded in Structures: Theory and Application to Molecular Cybernetics. *Foundations of Science* [Online], 27(4), pp.1327–1345. https://doi.org/10.1007/s10699-022-09830-8.

Burgin, M. and Mikkilineni, R., 2022. Is Information Physical and Does It Have Mass? *Information* [Online], 13(11), p.540. https://doi.org/10.3390/info13110540.

Heller, M., 2019. How is philosophy in science possible? *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* [Online], (66), pp.231–249. Available at: <https://zfn.edu.pl/index.php/zfn/article/view/482> [visited on 6 October 2021].

Krzanowski, R., 2017. Minimal Information Structural Realism. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* [Online], (63), pp.59–75. Available at: <http://www.zfn.edu.pl/index.php/zfn/article/view/396> [visited on 16 May 2018].

Krzanowski, R., 2020. Why can information not be defined as being purely epistemic? *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* [Online], (68), pp.37–62. Available at: <https://zfn.edu.pl/index.php/zfn/article/view/494> [visited on 30 November 2021].

Krzanowski, R., 2022. *Ontological Information: Information in the Physical World* [Online]. Vol. 13, *World Scientific series in information studies*. Hackensack, NJ: World Scientific. https://doi.org/10.1142/12601.

Mścisławski, Ł., 2022. Is information ontological or physical, or is it perhaps something else? some remarks on Krzanowski's approach to the concept of informationn. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (73), pp.147–169.

Polak, P., 2019. Philosophy in science: A name with a long intellectual tradition. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* [Online], (66), pp.251–270. Available at: <https://zfn.edu.pl/index.php/zfn/article/view/472> [visited on 6 October 2021].

Polak, P. and Trombik, K., 2022. The Kraków School of Philosophy in Science: Profiting from Two Traditions. *Edukacja Filozoficzna* [Online], (2(74)). https://doi.org/10.14394/edufil.2022.0023.

Shannon, C.E., 1949. *The mathematical theory of communication*. Urbana: University of Illinois Press.

Shannon, C.E. and Weaver, W., 1964. *The Mathematical Theory of Communication*, *Illini Books*, IB-13. Urbana: University of Illinois Press.

Turek, K., 1978. Filozoficzne aspekty pojęcia informacji. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (1), pp.32–41.

Turek, K., 1981. Rozważania o pojęciu struktury. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (3), pp.73–95.

Why is neuron modeling of particular philosophical interest?

Paweł Polak
Pontifical University of John Paul II in Kraków, Poland

Andrzej Bielecki, *Models of Neurons and Perceptrons: Selected Problems and Challenges*, Studies in Computational Intelligence, vol. 770, Springer International Publishing, Cham 2019, pp. 156.

A peculiarity of "philosophy in science" (see Heller, 2019; Polak, 2019) is that the best sources tend to be atypical from the viewpoint of most philosophers: For example, on the one hand, there are works that popularize science, while on the other hand, there are research articles and even specialized monographs. The book discussed here falls into the last category, and it is devoted to modeling neurons and perceptrons. It was written by a mathematician from Kraków, Andrzej Bielecki, who is currently working at the AGH University of Science and Technology.[1] Readers of *Philosophical Problems in Science/Zagadnienia Filozoficzne w Nauce* (ZFN), as well as the related *Semina Scientiarum* journal, will probably associate him with the philosophical activities that he has practiced within the context of his scientific activities (2016; 2018). Bielecki is an example of a scien-

[1] Andrzej Bielecki received an M.Sc. degree in Physics and Mathematics and a Ph.D. in Mathematics, D.Sc. (habilitation) in Mathematics from the Jagiellonian University in Kraków. He obtained a professorship in Computer Science in 2020. His fields of interest includes dynamical systems theory, artificial intelligence, cybernetics, and philosophy of science, and he has written over 120 scientific papers and one textbook.

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 347–356

tist–philosopher from the Krakow milieu,[2] and it is worth noting that he develops his philosophical activities, among other things, through his work on the Committee on Philosophy of Science at the Polish Academy of Arts and Sciences in Kraków.

Bielecki's book is published as a volume in the "Studies in Computational Intelligence" series, which is intended for research that contributes to computational intelligence. The book is an in-depth monograph about computationally modeling basic cognitive structures, such as neurons. It comprises five parts that logically present different areas of the subject. The first part, titled "Preliminaries," provides fundamental biological knowledge about neurons and essential information about the basics of artificial neural networks and their applications. The second part is then devoted to the mathematical foundations of modeling, particularly dynamical systems theory. Next, the third part goes into mathematical models of neurons, such as models of entire neurons and models of portions of neurons. The fourth part then focuses on modeling perceptrons, starting with linear perceptrons and ending with nonlinear ones. The final part consists of the appendices.

The author deliberately combines biological and simulation perspectives in his book, aspects that are usually separated into distinct studies within neuron research. This interdisciplinary approach aims to identify new sources of biological inspiration for mathematics and computational modeling. Bielecki also says he chose this approach "because it seems that there are numerous models of biological neural structures that can be the basis for artificial systems and that have not been utilized yet" (Bielecki, 2019, p.3). It is worth adding here

[2] It is worth mentioning that in the book, Bielecki mostly uses examples from research conducted in the Kraków milieu. The use of the cybernetic theory framework could also be interpreted as another sign of the local milieu's influence.

that although it is not explicitly stated in the book, Bielecki's attitude toward interdisciplinarity has resulted from his work in various interdisciplinary research teams that have included biologists and mathematicians.

Bielecki's work provides an essential overview of the contemporary view of neuron modeling, and the included bibliography is a helpful further guide in this area. Here, the reader can find extensive and detailed, yet concisely presented, knowledge about modeling neurons and their networks. By zooming in on this monograph's detailed explanation of the problems of modeling a single neuron, we can quickly realize how simplistic assumptions are often made in projects related to Whole Brain Emulation (WBE). For my part, I regard this as a warning to approach the results of WBE-like projects with extreme caution (e.g. Kycia, 2021). After all, a single neuron itself is still not sufficiently understood (e.g. Bielecki, 2019, p.133), and the complexity of its structure leads one to realize the incredible complexity of the brain, as well as the level of complexity we are trying to master in brain-related research. Even the problem of practical computational complexity in whole-neuron modeling comes up: "It should be stressed that, currently, the computational power of computers is too weak to compose the model of the whole neuron by using models of its parts" (Bielecki, 2019, p.59). The author also notes the need for inter-level studies (i.e., between subneural, neural, and network levels). We should add that if we talk about the emergence of properties at higher levels in philosophy, such topics are consistently overlooked in scientific research.

Bielecki's monograph makes one realize how much effort we should be devoting to discussing the role of simplifying assumptions and idealizations in simulations. Of course, artificial neural networks (ANNs) can be based on greatly simplified models of neurons for

technical applications and often succeed at achieving the desired goals. The situation is different in scientific research, however, because it attempts to describe the functioning of neuronal structures, such as the brain, through simulations. Bielecki states this clearly: "In the light of neurophysiological knowledge, the models of the whole neuron are simplified to such an extent that they do not reflect, even approximately, the character of signal processing in the biological neuron." (Bielecki, 2019, p.57)

It is worth noting that a particular strength in Bielecki's book is how he does not limit modeling to just standard computer modeling. Indeed, he is also interested in physical (electronic) models that operate on continuous values due to the problems with digital simulations of nonlinear differential equations: "If the model is based on ordinary differential equations, then it can be implemented by using an electronic circuit whose dynamics is described by the same differential equation" (Bielecki, 2019, p.17). Bielecki proposes using a kind of classical analog computation. From a philosophical perspective, this means he does not share the common tacit philosophical assumption among many works that Turing's computational model can sufficiently describe the real world. For this reason alone, I think that any philosopher who wishes to make a responsible statement about neurons, the brain, and the research about them should read this book. Reflecting on the implications of the knowledge presented here should help the reader to understand how many problematic assumptions we currently make in discussions related to this topic. I would like to share some of my thoughts that were inspired by this book below.

The reviewed monograph brings some exciting contributions to the discussion about simulation methodology in biology. Indeed, the specific issues of biological simulation are worthy of a separate study, which, by the way, Bielecki is currently working on. Nevertheless, the

methodological specifics of such simulations are rarely addressed. In Bielecki's book, however, we can find an attractive methodological scheme that has the advantage of being created based on scientific practice. It is therefore an excellent example of "philosophy in science," which in this case is located at the intersection of applied mathematics and biology.

In Bielecki's view, computational modeling begins with biological research (A), which allows us to distinguish relevant structures and processes. The next step then requires biological experiments or observations (B). The crucial properties can only then be determined (C) based on these, enabling a semi-formal description (D) to be formulated. This description can then serve as the basis for creating a formal model (E), which can then act as the basis for constructing a software or hardware implementation (F). According to Bielecki, these final two stages can influence each other, with each acting as the starting point for formulating the other. Finally, it is essential that the results of formal modeling should eventually become the subject of an analysis through a traditional approach (G). Consequently, it may become necessary to modify the experimentation/observation phase (B) or the determination of the crucial properties (C). Such feedback is essential to the computational modeling methodology, but it also indicates how much creative input the scientist has. Models are not mere generalizations of facts, as methodologists once wanted them to be, but rather the result of a complex, looped adaptive process.

Interestingly, the precondition for creating such models—and therefore the need for learning more about complex, or perhaps *more* complex, biological structures—is the ability to perform sufficiently complex calculations, either in digital or analog form. The methodological scheme indicated by Bielecki therefore points to a strongly "non-linear" looped process that occurs during the creation of ad-

vanced biological knowledge. It is a case of epistemic bootstrapping, or more precisely, it could be described as epistemic feedback (Weisberg, 2010). Interestingly, an essential argument for considering such a "non-linear logic of scientific development" (to use Heller's words) flows directly from scientific practice. Bielecki, however, is not interested in isolated arguments for and against epistemic bootstrapping. He instead posits the validity of this method based on an analysis of actual scientific practice in biology. It should be emphasized that the reviewed book does not contain detailed philosophical analysis or present a pro and contra discussion of the presented theses but rather seeks to uncover an essential philosophical issue that is entangled with modeling in biology. Nevertheless, meticulous analyses and deliberations about the pros and cons should be the next step in reflecting upon the philosophical issues of biological simulations. Nevertheless, let us highlight that such an endeavor would not be possible without first identifying these issues, and this book plays an important intellectual role by posing important and non-trivial philosophical questions, even if it does so indirectly.

Bielecki's monograph also shows the level of depth in the mathematization of biology that is taking place in research at the cellular and subcellular levels. The author does not apply the slightest hint of persuasion here but rather simply demonstrates the impressiveness of the precise mathematical basis for neuronal modeling. It easily convinces the reader of the deep and practical mathematization of biology that has taken place through computational modeling and the adoption of a cybernetic framework.

Bielecki's remarks about the need to synthesize various modeling approaches are worth special attention: "In this monograph, the cybernetic modeling, the mathematical modeling, and the modeling by using electronic circuits intertwine. [. . .] This is also a specificity of

the approach presented in this monograph because these three ways of modeling are usually exploited separately." He also points out this approach's more general, philosophical context: "Since the Enlightenment analytic approach to scientific problems has dominated, and the synthetic approach is, in general, in the state of atrophy. The synthetic mathematical–electronic approach to modeling sub-neural processes, presented in this monograph, tests whether such an approach can be efficient. *The results show that the answer is affirmative* [emphasis added]" (Bielecki, 2019, p.124). Note that I emphasized the final sentence to highlight how the author sees this book as a kind of methodological experiment with a positive result. Indeed, I think this result should be presented to philosophers in more detail to help us understand its methodological soundness, and maybe a separate study on this issue could be appropriate for clarifying Bielecki's ideas.

Now, let me illustrate the conceptual scheme used by Bielecki: It is based on concepts from cybernetics theory, one of the vital mathematical theories that provides the foundation for developing interdisciplinary research and computational modeling. In Poland, cybernetics is still successfully pursued, especially in Kraków at the AGH University of Science and Technology,[3] but contemporary international discussions use somewhat different conceptual systems. A good example is Gordana Dodig-Crnkovic's article in an issue of ZFN (Dodig-Crnkovic, 2022). The deep analogies between the two approaches are surprising. For example, take Bielecki's phrase: "Each type of biological cells, including the simplest bacteria, receives stim-

[3] In private correspondence, the author stated that the most important sources of inspiration on the issue of cybernetics are the works of Tadeusiewicz (1994; 2009), who is a distinguished researcher and the founder of a vivid center of biocybernetics at AGH in Krakow. A further source of inspiration were the works of another Krakow scientist, Mariusz Flasiński (1997; 2016), who is affiliated with Jagiellonian University in Kraków.

uli from its environment and processes the obtained signals" (Bielecki, 2019, p.5). It is close to the info-computational in Dodig-Crnkovic's view, although she uses a specific reference to information theory. It would be worthwhile to analyze the relationship between cybernetics and contemporary information concepts in more depth, because it may be possible to find new, inspiring analogies or more convenient conceptual frameworks.

Finally, let us conclude with the specifics of "philosophy in science," with which I began this review. One of its unique features is that interesting contributions can be rich in philosophical content, even though the word "philosophy" may rarely appear in them, if at all. Andrzej Bielecki's book is an excellent example of this, because he mentions philosophy only twice, and one of those refers to the Enlightenment. Nevertheless, it makes an exciting contribution to understanding the philosophical issues in modern biology.

Abstract

This review article discusses Andrzej Bielecki's book *Models of Neurons and Perceptrons: Selected Problems and Challenges*, as published by Springer International Publishing. This work exemplifies "philosophy in science" by adopting a broad, multidisciplinary perspective for the issues related to the simulation of neurons and neural networks, and the author has addressed many of the important philosophical assumptions that are entangled in this area of modeling. Bielecki also raises several important methodological issues about modeling. This book is recommended for any philosophers who wish to learn more about the current state of neural modeling and find inspiration for a deeper philosophical reflection on the subject.

Keywords

neuron modeling, sub-neuron modeling, computational modeling, analog computation, philosophy in science, philosophy of biology, philosophy of computing.

Bibliography

Bielecki, A., 2016. Cybernetyczna analiza zjawiska życia. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* [Online], (61), pp.133–164. Available at: <https://zfn.edu.pl/index.php/zfn/article/view/361> [visited on 27 January 2020].

Bielecki, A., 2018. Epistemologiczne problemy w biologii subkomórkowej: obserwacje, modele matematyczne i symulacje komputerowe. *Semina Scientiarum* [Online], 16, pp.10–23. https://doi.org/10.15633/ss.2482.

Bielecki, A., 2019. *Models of Neurons and Perceptrons: Selected Problems and Challenges* [Online]. Vol. 770, *Studies in Computational Intelligence*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-90140-4.

Dodig-Crnkovic, G., 2022. In search of common, information-processing, agency-based framework for anthropogenic, biogenic, and abiotic cognition and intelligence. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (73), pp.17–46.

Flasiński, M., 1997. "Every Man in His Notions" or Alchemists' Discussion on Artificial Intelligence. *Foundations of Science* [Online], 2(1), pp.107–121. https://doi.org/10.1023/A:1009687513096.

Flasiński, M., 2016. *Introduction to Artificial Intelligence* [Online]. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-40022-8.

Heller, M., 2019. How is philosophy in science possible? *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* [Online], (66), pp.231–249. Available at: <https://zfn.edu.pl/index.php/zfn/article/view/482> [visited on 6 October 2021].

Kycia, R., 2021. Information and brain. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* [Online], (70), pp.45–72. Available at: <https://zfn.edu.pl/index.php/zfn/article/view/514> [visited on 6 December 2022].

Polak, P., 2019. Philosophy in science: A name with a long intellectual tradition. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* [Online], (66), pp.251–270. Available at: <https://zfn.edu.pl/index.php/zfn/article/view/472> [visited on 6 October 2021].

Tadeusiewicz, R., 1994. *Problemy biocybernetyki*. Wyd. 2. Warszawa: Wydawnictwo Naukowe PWN.

Tadeusiewicz, R., ed., 2009. *Neurocybernetyka teoretyczna*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Weisberg, J., 2010. Bootstrapping in General. *Philosophy and Phenomenological Research* [Online], 81(3), pp.525–548. Available at: <https://www.jstor.org/stable/41057492> [visited on 6 December 2022].

Is AI case that is explainable, intelligible or hopeless?

Łukasz Mścisławski
Wrocław University of Science and Technology, Poland

H. Cappelen, J. Dever, *Making AI Intelligible: Philosophical Foundations*, Oxford University Press, Oxford 2021, 175+viii pp.

The bored reader may sigh: another book in which philosophers create an artificial problem without completely understanding artificial intelligence (AI). It is entirely legitimate here to ask whether philosophers' throwing themselves (and—to be honest—not only them) at various AI problems is fruitful and the subject matter is really important? Does AI itself, in order to be somehow tamed (to function in a more or less communicative way with us)—need such a far-reaching intellectual effort, by not only technical in its nature? Some justification can be found in the surprising effectiveness of AI systems in relation to the tasks set before them, which arouses understandable interest and is sometimes heavily exploited in the media. Nevertheless, some revolutions of extreme importance, involving digital systems, as Paweł Polak (2015, p.151) rightly pointed out, proceed without special publicity. Perhaps, at least to a large extent, this would also be the fate of AI systems if it were not for the fact that decisions can be made based on them about important issues in the lives of ordinary people.

Herman Cappelen and Josh Dever, in their book *Making AI Intelligible: Philosophical Foundations*, provide a positive answer to both questions. The scope of the subject matter addressed in the book is

Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)
No 73 (2022), pp. 357–369

quite narrow and is mainly concerned with the issue of the possibility of linking the content on which humans can operate with the way AI systems function and deliver results. The authors raise a number of important issues that become more pressing as AI systems penetrate more and more new areas of human functioning. It also turns out that attempting to theoretically justify the answers to the questions that arise is far from easy, despite the existence of an extremely rich set of different philosophical traditions, equipped with powerful tools developed to solve various problems.

Chapter 1 (*Introduction*) presents the primary task of the book: an attempt to answer the question of whether philosophical theories of meaning, content and language can be helpful in understanding, explaining and—perhaps—improving AI systems?

According to the authors, the answer to the question posed in this way is positive. They begin their argument by presenting a situation in which a decision concerning a fictitious person is made by an AI system. The question is about the possibility of granting credit. The answer is negative. This raises a simple question: why? The authors point out that knowing how AI systems work does not directly translate into understanding the results provided by such systems. Much worse, however, is the attempt to reconstruct what is the rationale for such and not such a result (pp.4–10).[1] For all the effectiveness of such tools, the fundamental difficulty, on the unraveling of which the authors devote practically the entire book, lies in the fact that it is not very clear whether and how the information processing processes taking place in AI systems are related to the content on which humans can operate. Attempting to answer this question can be seen as a waste of time and the answer itself as adding little—at least from the point

[1] All page numbers without Author and year refer by default to (Cappelen and Dever, 2021).

of view of those designing and implementing such systems. This kind of working conclusion seems to conclude the exemplary dialogue with a hypothetical AI specialist, to which the whole of Chapter 2 is devoted (*Alfred (The Dismissive Sceptic). Philosophers, Go Away*!). Nevertheless, already in Chapter 1, the authors make important points: the need for a good content theory for AI systems and the pressing issue of relating the output of such systems, expressed through a group of sentences in some natural language, to the content determined in that language by that very group of sentences (pp.20–21). However, the answer to the question of why to explore such a seemingly insignificant issue turns out to be very important, given the increasingly widespread use of AI-based decision-making systems. Although no such statement is made explicitly, an attempt to reconstruct such an answer from an already preliminarily sketched, fictional dialogue between a philosopher and a specialist about AI could be as follows: the link between the results provided by AI, their justification and the content on which humans operate is important, as AI systems are increasingly being incorporated into decisions concerning existential human affairs. These include the possibility (or not) of e.g. taking out a loan, health matters, but also making a diagnosis or the adjudication of being a criminal suspect (pp.36–38). Hence, the suggestion that reliance on these systems simply because it is well-written software, based on sophisticated mathematical apparatus, seems insufficient at best (p.37). It should be emphasised that the authors are not concerned with some kind of embedded content in AI systems. Rather, they are concerned with understanding what the content is in a given complex system and how that content was obtained by that system (pp.22–23). This is particularly true for AI systems, the results they provide and their interpretation (p.18). Here is right place to highlight is one of the minor shortcomings of the work. The aforementioned thesis, that

the link between the results provided by AI, their justification and the content on which humans operate is important, as AI systems are increasingly being incorporated into decisions concerning existential human affairs and the suggestion that reliance on these systems only because it is well-written software, based on sophisticated mathematical apparatus, is not formulated explicitly. Moreover, with regard to AI-based systems, the issue of trusting the software is one thing, but there are also other problems: the problem of AI bias (which luckily is addressed by Author, however rather in technical context and with reference to content issues), the issue of quality and ethics and the value system used in AI training (e.g. Spence, 2021) and the fundamental question of the correctness of the mathematical model (and its adequacy to the simulated area of reality). Although the last issue is not the direct focus of the authors' research, it seems that some mention of such difficulties would be most welcome.

Chapter 3 (*Terminology*) is devoted to introducing basic concepts, which is important for the clarity of the overall discussion and introduces the reader to aboutness, representation, and attempts to outline the connections between these concepts and AI, metasemantics and philosophy of mind. Building on a previous dialogue with a sceptic, the authors note that, in essence, software or devices in themselves say nothing. The analogy is with AI systems. Moreover, an attempt to build an understanding of the results provided by AI systems, based on knowledge of their internal structure and operating principles, does not necessarily shed much light here. This, in turn, leads to the conclusion that there is a need for a stronger interplay between the metaphysics of content and theories of AI, and a suggestion to look more closely at the possibility of using the tools provided by the

externalist branch in the philosophy of language. The authors see the lack of a wider discussion of this possibility in the literature as a gap that needs to be filled (pp.53–58; cf. Krzanowski and Polak, 2022).

In Chapter 4 (*Our Theory. De-Anthropocentrized Externalism*), the authors make a presentation of their own conception, which they describe as de-anthopocentrised externalism. They base their proposal on two basic claims: 1) the content of AI systems should be explained externalistically; and 2) existing externalist approaches are anthropocentric. The first thesis is based on the observation that content, related to action, is not a problem at the level of software or computation. It is a problem at the environmental and sociological level. The second thesis is the observation that however philosophers have developed impressive models of human language and human mental states. However, this is not the case with AI systems—the operation of software on specific hardware, both in the computational layer and in terms of the functioning of the hardware, is fundamentally different from what can be described by such means. A de-anthropocentised metasemantics is therefore needed here (pp.59–71). However, some additional rule of thumb is also needed for the future selection of appropriate measures and the development of an effective content theory of AI systems. Here, the authors propose a meta-metasemantic principle: interpreter-centric knowledge-maximization. Two important issues also arise here, which can also serve as a kind of guideline in the search for appropriate tools for further research: a) it is the human knowledge and not the AI system that is important, so the idea is to maximise human knowledge; and b) the perspective of interests is important here, bearing in mind that human interests may differ from those of the an AI system[2] (pp.75–79).

[2] In simple terms: a human may want to know why he or she was classified negatively in given aspect, while an AI system may seek to optimise the data in some way (e.g. finding the minimum of a function).

The background thus outlined serves the authors to attempt to apply already existing philosophical tools when it comes to relating content to the results provided by AI systems. The test task is to classify a particular person into a particular category. The basic problematic therefore involves AK1) referring to that particular person; AK2) attributing to that person a test characteristic (adjudicating that person as having that characteristic); AK3) criteria for classifying that person into a particular category (adjudicating attribution to a category). Chapter 5 (*Application. The Predicate 'High Risk*') attempts to unravel these issues using an externalist approach based on proposals of Kripke. In doing so, they draw attention to the fundamental difficulties of such approaches: the problem of the anchoring event, the problem of defining chain of transmission and issuses connected with the problem of being part of communicative chain, when AI systems are involved (p.82-88). Additional difficulties are posed in relation to AK3 by the possible variability of classifications and models and the fact of context dependence. Unsurprisingly, it is very important to note that in the case of systems based on machine learning, the final correctness of the answers given depends on those that the human training such systems deems to be true, i.e. on human decisions. Another problem is that AI systems do not seem to have capacity to representing that could be analogous to human's ability of representing using proper names. Such a situation generates serious problems of communicative and epistemic nature (pp.99–105). Cappalen and Dever make an analogous analysis when it comes to the potential application of the Mental Files Framework[3] and attempt to extend the

[3] Murez and Recanti shall be characterized as *devices of direct reference whose deployment makes it possible to entertain singular thoughts, i.e. thoughts that are about particular objects rather than about whatever possesses certain features or satisfies such and such a description* (cf. Murez and Recanati, 2016, p.267).

findings of Evans and Recanati to Epistemically Rewarding Relations. Chapter 6 (*Application. Names and the Mental Files Framework*) is devoted to this. The addition of the knowledge-maximization rule to the framework in question raises the question of whether it is about maximising general or specific knowledge (p.114). Ultimately, however, it appears that with the philosophical tool in question the case is similar to that of the Kripke-style framework, a simple application to AI systems may not be feasible. In particular, there is need to abstract the notion of an epistemically rewarding relationship. The main difficulties in the context of the philosophical framework in question, however, are the need to focus on particular epistemic goals and activities and the fact that what the results provided by an AI system are about depends on the aims of the interpreter. Hence the results of the considerations in Chapters 4 and 5 are reinforced and shows the organic nature of the internalist view of AI: you cannot bite off all the facts about the classification content of a machine learning system by looking only at the internal implementation of that system (pp.115–116).

It should be emphasized that the analyses carried out of the positions presented in the chapters under discussion are very detailed and thorough. The bigger surprise is Chapter 7 (*Application. Predication and Commitment*), which turns out to be a fundamental twist. The authors conclude that the attempts presented earlier to link human-understandable content to the way AI systems function are far from sufficient. They therefore pose the thesis that this kind of linkage is not only a denotation of something, but also an act of adjudication. They thus introduce the reader to the foundations of the Act Theoretic View, which seems a legitimate step insofar as, following Soames and Hanks, they assume that propositions do not have intrinsic representational properties. This in turn—at least to some extent—gets rid

of the problems of semantic externalism. After a brief introduction, the whole chapter is essentially devoted to an attempt to investigate whether it is possible to use such a tool to solve the problem of interpreting the performance of AI systems and the results they provide. (pp.117–121). The metasemantic tool that Cappalen and Dever choose to use in relation to predication is the Teleofunctionalist Hypothesis (TFH). They formulate TFH as the statement that a mental act is the act of predication because of its teleofunctional role in giving rise to judgements that guide action. Using proposed tools gives them also possibility to not committ to any particular architecture of analyzed system (pp.123–125).

In doing so, they point out that the TFH approach also presents peculiar difficulties. An example of this is the changing objectives that ultimately help to provide an answer with a given content.[4] However, the aforementioned independence from specific AI system architectures should be considered a very strong advantage. They also propose to consider the relationship between TFH and commitment (or assertion) and try to infer some some norms that could be results of theory based on such approaches. At the end of the chapter, they propose an outline for such research project that could attempt to explore theories of assertion and commitment for humans and AI. Although one of the authors (Cappalen) is sceptical about the category of assertion itself, for the inquisitive reader this outline will undoubtedly provide some inspiration for their own research.

The last chapter (*Four Concluding Thoughts*) contains a kind of explication of the threads that, however scattered, appeared in the previous chapters. The first point is that AI systems are dynamic realities in the senescence that they have a kind of dynamic purpose. Such a situation requires a little more knowledge of technical details

[4] E.g. positive or negative classification for a mortgage.

on the part of philosophers, which, for example, a is essential when dealing with issues related to the scoring mechanism and the problem of the number of layers used for characterization of content (p.141, pp.140–148).

The second point is for the authors to consider the applicability of the philosophical description associated with 'active externalism', as proposed by Clark and Chalmers, and the concept of extended mind presented by them (pp.148–157). The problem is important because, as the authors point out, the effort to understand extraneous content, and the content contained in AI systems can be considered as such, turns into the issue of understanding the determination of content within our extended mind (p.156). A point to be made here, however, is that it seems to be one thing to determine content and another to understand what these extensions of the mind operate on and what is the relation of the extended mind as a whole to the content on which humans operate.

The third point is an attempt to completely change the position, which here Cappalen and Dever refer to as a content-driven approach. The authors sketch an attempt to justify an application to the issues considered in the book from the point of view of the No-Content-Just-Evidence approach. In doing so, however, they draw attention to the problems of Adversial Perturbations, ML system bias and the important fact that coincidental convergence is not justified enough to treat AI systems as reliable for new cases (pp.157–162). This raises the question of the justification of trust in AI systems. This brings the authors, at least in a sense, to the fourth point and some kind of connection with *Explainable AI* stream. It is appropriate to cite their objections to this stream: a) without content there are no reasons nor explanations, and AI systems 'says' something that is contentful. Also reasons are contentful themselves; b) very explainability is also

a process of determining specific content. Hence there is great need to say something about content and its connection with explanation in context of AI systems (pp.162–165).

What can be seen as very strong point of book under review is the observation that talking about issues referring to functioning of AI systems there are many anthropomorphism used. Meanwhile, in the case of AI, the matter is quite complicated. As Authors put it:

> In philosophy, consideration of alien languages either starts with the assumptions that the aliens share with us a basic cognitive architecture of beliefs, desires, reasons, and actions, or (as Davidson does) concludes that if the aliens aren't that much like us, then whatever they do simply can't count as a language. Our point is that the aliens are already among us, and they're much more alien than our idle contemplation of aliens would have led us to suspect. Not only that, but they are weirdly alien—we have built our own aliens, so they are simultaneously alien and familiar. (p.17)

This shows how difficult it is to connect more or less obvious for a man content of sentences used by his language with, as it seems, complete alien world of AI systems, of which technical structure and algorithms, paradoxically, we know almost all.

It should be emphasized that presentations of ideas and argumentations that lead to using externalistic metasemantics in interaction with AI, are very clear and is one of the strongest points of the book. The book has, however, also some shortcommings. They do not decrease the value of the work of Cappelen and Dever, nevertheless they are confusing and hinder reading of this fascinating book.

The issue of the relationship between human understanding of and operation on content and how AI systems function is, on the one hand, an extremely important task, but also a very difficult one. It

could be said that one of the weaknesses of the book is the lack of attempt to outline what the authors actually mean by the content. This could have helped the reader grasp more of what the difficulty of the whole endeavour is, above all in comparison with operating on, for example, colours within graphic systems.

Perhaps also devoting some space to other seemingly obvious issues would have helped not only the readers but also the authors in their endeavour. It would seem that already with regard to the aforementioned notion of content, it would at least be appropriate to outline overtly some ontological background within which the authors conduct their analyses. There would then be a chance for various hidden assumptions to see the light of day and this would give an opportunity to assess their impact on the overall argumentation carried out. One such assumption, by no means obvious, is to treat AI systems as designed and made intentionally, as opposed to human (p.70). However, that humans are not created intentionally seems to be a very strong ontological assumption.

The starting point for the strategy proposed by the authors, as could be seen in the brief presentation of the individual chapters, is an attempt to use attribution mechanisms that appear to be human-specific (which is the case in Chapter 5). However, the authors are aware that this kind of tactic cannot be applied across the board, as witnessed in Chapter 4, and the de-anthropocentric perspective proposed therein. However, when, as they rightly point out, this path does not yield satisfactory results, they change the tools with which they try to get their way (Chapter 6 and later Chapter 7). While this is understandable, it seems that the link between abandoning the previous path of trying to deal with the problem under investigation and the choice of subsequent tools is not sufficiently justified. A side effect of such a situation may be a feeling that the reading of the

book is piecewise smooth. The authors also do not make any remarks, whether the failure of the given tool in question is a permanent or whether it is only temporary situation and we need more research in given area or wait for more advanced technologies.[5]

The book may leave you feeling unsatisfied a little bit. At the moment when the narrative gains momentum, the book ends with few important and accurate remarks on explainable AI, but also is leaving the reader with only an outline of further possibilities for continuing the plot. However, this is quite understandable due to the fact that the issues raised by the authors are extremely extensive. Attempting to cover all possible approaches to the issues raised would fundamentally break the frame of any book of reasonable length. Nevertheless, in their work Cappelen and Dever is very inspiring, it poses many very important questions, tries to find solutions and provokes independent study.

Abstract

This article is a review of the book *Making AI Intelligible. Philosophical Foundations*, written by Herman Cappelen and Josh Dever, and published in 2021 by Oxford University Press. The authors of the reviewed book address the difficult issue of interpreting the results provided by AI systems and the links between human-specific content handling and the internal mechanisms of these systems. Considering the potential usefulness of various frameworks developed in philosophy to solve the problem, they conduct a thorough analysis of a wide spectrum of them, from the use of Saul Kripke's work to a critical analysis of the explainable AI current.

[5] The answer to that question seems to be particularly in case of explainable AI.

Keywords

AI, externalism, metasemantics, content.

Bibliography

Cappelen, H. and Dever, J., 2021. *Making AI Intelligible: Philosophical Foundations* [Online]. 1st ed. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780192894724.001.0001.

Krzanowski, R. and Polak, P., 2022. The Meta-Ontology of AI systems with Human-Level Intelligence. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (73), pp.197–230.

Murez, M. and Recanati, F., 2016. Mental Files: an Introduction. *Review of Philosophy and Psychology* [Online], 7(2), pp.265–281. https://doi.org/10.1007/s13164-016-0314-3.

Polak, P., 2015. Bezgłośna komputerowa rewolucja w naukach eksperymentalnych [recenzja]. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, (58), pp.151–157.

Spence, E., 2021. *Stoic Philosophy and the Control Problem of AI Technology: Caught in the Web*, *Values and identities*. Lanham: Rowman & Littlefield Publishers.