# FILOZOFIA I NAUKA

Studia filozoficzne i interdyscyplinarne

**Tom 10, zeszyt specjalny** **2022**

Instytut Filozofii i Socjologii
Polskiej Akademii Nauk
Instytut Filozofii
Uniwersytetu Marii Curie-Skłodowskiej

ISSN 2300-4711
ISSN 2545-1936 (ONLINE)

## RADA REDAKCYJNA

# FILOZOFIA I NAUKA

## Studia filozoficzne i interdyscyplinarne

**Tom 10 , zeszyt specjalny, 2022**

**Instytut Filozofii i Socjologii Polskiej Akademii Nauk**


**Instytut Filozofii Uniwersytetu Marii Curie-Skłodowskiej**

**Kontakty**

Adres redakcji: Instytut Filozofii i Socjologii Polskiej Akademii Nauk,
pokój 104, ul. Nowy Świat 72, 00-330 Warszawa
Numer telefonu: 603 160 505
adres elektroniczny: filozofiainauka@ifispan.edu.pl
Strona internetowa: www.filozofiainauka.ifispan.edu.pl

**Dostęp**

Egzemplarze wersji papierowej można kupić, cena – 20 złotych. Zamówienie z adresem należy składać na elektroniczny adres redakcji.
Czasopismo w wersji elektronicznej jest wydawane w trybie open access: streszczenia wszystkich opublikowanych w czasopiśmie tekstów są dostępne na stronie internetowej czasopisma. Pełne teksty są udostępniane 1) na stronie internetowej czasopisma równocześnie z wersją papierową, 2) na platformie EBSCO, 3) na platformie cyfrowej PAN – Czytelnia Czasopism, a abstrakty – są publikowane w CEJSH (The Central European Journal of Social Sciences and Humanities).

Drukarnia: Paper & Tinta, Warszawa

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

# PHILOSOPHY AND COMPUTING: AI, VIRTUALITY, EPISTEMICITY

Guest-editor: *Piotr Bołtuć*

TABLE OF CONTENTS

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

## *Editorial*

# PHILOSOPHY AND COMPUTING: AI, VIRTUALITY, EPITEMICITY

doi: 10.37240/FiN.2022.10.zs.0

### 1. THE CALLING OF PHILOSOPHY AND COMPUTING TODAY

We need a productive, veridical narrative related to the current growth of artificial intelligence and its social role.

The narrative of fear that came from the early movies and novels is great for a thriller; but in terms of social discourse, it puts us at the level of Russian peasants, during the reign Tzar Alexander, kneeling at the view of the early trains, taking them to be a deed of the Satan.

The narrative of dismissal, claiming that those are just tools and need to be treated as such, is highly outdated since computer programs and robots are much more already, and their capabilities increase. The approach of Isaac Assimov, prescribing that we must view computers as slaves, limited to obedient following of human orders, came from the narrative of dismissal, and turned it into a narrative of enslavement of intelligent beings such as AI. *My criticism of Assimov's approach is not particularly guided by moral disagreement with him (at this point moral status of AI is unclear, and needs much further work and development of future AI before it gets settled). It is based on practical considerations. It is a waste to treat a philosopher as a slave (to use Plato's example) since he would be a terrible slave, and if there is social use for philosophers it is clearly not facilitated through enslaving them. The same goes for advanced AI—it it just waits for us to decide, the car it "helps drive" is likely to crush and the armed rocket it can strike down is going to reach its target before its human handler can figure out what the matter is. We are just not efficient enough to be in charge of those new generations of smart and efficient being—at least at the level of fast implementation.*

The narrative of surrender is also a bad choice, just like excessive fear or dismissal. From my criticism of those attitudes, it does not follow that AI should become in charge of the human world. In extremally time-sensitive

predicaments, and in complex multi-factor mathematical analysis—human beings cannot even conceptualize some of the factors that surface through advanced big data analysis within open conceptual frameworks. We should be able to set up the general goals, exclude and include some of the acceptable means of attaining them (e.g. through including moral and high level legal imperatives, as well as the economic and other objectives).

Thus, *the narrative of optimal balance between human and artificial intelligence* emerges as the sole strategy to move forward, towards human flourishing, not engrossing oneself into the *triller-quality pessimism* or some tendencies to reverse towards the slavery economy—now by oppressing not only the enslaved human beings[1] but also artificial intelligence.

The general drift of the Asian philosophy and economy, for instance in China and Japan (however different those mentalities and social practices are), makes it easier to conceptualize such bi-directional approach than getting it through o of the other of conceptual frameworks. With all the creative potential of Western cultures, it is worth trying to develop a closer understanding of the mechanics and semiotics of the world with the artificial beings playing a role of much more than the tools, but effectively balanced by the human good, values and objectives.

Philosophy of Computers and especially Philosophy of AI plays the role of carrying this relevant debate beyond the academic lecture-room or strategy think tanks among the politicians or business leaders.

In the current issue of *Filozofia i Nauka* [Philosophy and Science] some of those topics are posed directly; but most are conversation openers for further debates. The more various routes we explore the better the chances of a reaching constructive world-views in the epoch of AI.

## 2. PHILOSOPHY AND COMPUTING TODAY

This issue of *Filozofia i Nauka* presents some of the important aspects of Philosophy of Computing in 2021–2022, which is when all those papers have been created. The first part of the volume is devoted primarily to what we decided to call: Philosophy shaped by Artificial Intelligence (AI), or artificial general intelligence (AGI). It is essential since AI, especially those projects that pave the way towards AGI, open substantial philosophical issues; the rapid growth of those domains makes those issues even more relevant at this very moment.

The second area is Virtual Space, which becomes more and more relevant for our daily lives. The move towards placing so much of our work and life activities online, which is a substantial aspect of the Economy 4.0 (Rogers,

[1] It seems like the society enslaving some smart beings would enslave more of them, including the many human beings—which is not only bad ethics but also bad economics.

2016), accelerated largely due to the social distancing of the COVID pandemics. The third area is a bit more amorphous; it centers around the topics such as perception, imagination, motivation and the mind, as understood through the prism of not only human or animal cognitive activities but also those of computers.

## 3. PHILOSOPHY SHAPED BY AI

This issue starts with an article by Mark Burgin and Rao Mikkilineni on the seven layers of computation. The authors go beyond the standard distinction between symbolic and sub-symbolic computing. They introduce, or revamp, categories such as *super-symbolic computation, hybrid computation, fused computation, blended computation, and symbiotic computation*. The article skillfully incorporates the background in philosophical semiotics, especially the works of Charles Sanders Peirce, which creates the leitmotif, which comes back a few times but becomes the dominant theme in the article by Ricardo Gudwin and Eduardo Camargo, which closes this section.

I wrote on non-reductive physicalism for advanced AI. What may be of interest is that *I identify the stream of awareness with what neuroscience calls creature consciousness*. This allows us to ditch substance dualism (few people view creature consciousness as a *non-materialist* substance) and take non-reductive physicalism seriously enough. Kyrtin Arteides follows with his proposal of Philosophy 2:0. The author argues that super intelligent AI systems may help human philosophers sort out their disagreements and check some of their ideas against the background of the current sciences. What may be fascinating, or refreshingly, in this intriguing article, is that the author writes from the viewpoint of an expert in AI consciousness, viewing philosophy in this framework. Arteides proposes "perspective maps" to maintain contextuality of knowledge (Figure 1), but then he wants to resolve the differences through a couple of rounds of *mediator feedback* (Figure 2). I happen to believe, that Arteides' approach, especially multi-core analysis (Figure 4), could be repeated in not so distant future, in the context of sub-symbolic, fused and other systems based *through and through* on fuzzy logic, thus able to capture blobs of meaning, instead of packing philosophical ideas in analytical semantics of sorts. Arteides' fascinating ideas related to philosophy as ecology of thoughts (Figures 6–7, 9), seem to be a step in a similar direction. This superb article, is still built on the assumption that philosophies must be *human-readable*, which seems like a non-controversial assumption, does not it?

The section closes with two rather thorough articles. Building on predictive coding and predictive processing, Jeffrey White explore the possibility of creating metaphysical self. This is a paper where boldness in philosophi-

cal thinking comes from interpretations of some work in AI. Most philosophers would view this approach as highly suspect; yet, taking into account the new ontologies created by various cognitive architectures in AI, many things become an open question. The article by Eduardo Camargo, Ricardo Gudwin on grounding cognition in Peircean semantics is a part of the fascinating project developed by the second author for several years now. The paper seems to keep just the right balance (if there is such a thing) between philosophy and AI, building at the conceptual space that belongs to both domains. We encounter a rather detailed in presentation of Peirce's semiotics (in section 3), then generalized in knowledge generation for both human and robot domains (section 4). In section 5 the authors focus on post-Peircean ontology that follows smoothly from the preceding arguments. Section 6 is devoted to the area of action and creativity, also largely based on Peirce's theory, yet updated and really focused on contemporary uses, which is advanced in section 7.

## 4. VIRTUAL SPACE IN PHILOOPHY

The article by Mariusz Mazurek devoted to ontology of virtual objects, provides a bridge between ontological reflection from Part 1 of the issue and the issue of virtuality dominating Part 2 The paper focuses on modes of virtual existence—from private objects on someone's screen through the process of their social objectifying. This results in intersubjective and often autonomous objects. The paper also presents important works in ontology of virtual objects, including those by Michael Heim, Jeri Fink, Lynn Baker and leading Polish authors.

This is followed by two articles from the 4th space group, whose agenda is to work our specificity of the virtual space in modern world. Bogdan Popoveniuc presents the interrelationship between personal and moral identity in the virtual space. This is probably the only "Continental" article in this issue, and Continental in a good sense it is, especially as phenomenology of the 4th space. The paper focuses on ontologies of the virtual space viewed both in philosophical and engineering perspectives, mostly as a technological extension of reality. It also leads to epistemological reflection, that bridges Part 2 with Part 3 of the current issue. Popoveniuc analyses space, including prominently the virtual space, as the foundation of subjectivity, engaging in deep reflection on philosophy of self. He analyses virtualization as a gradual cultural and technological process. The paper is a source of eruditional information from Gilles Deleuze and Félix Guattari on rhizomes and reality as maps (so compatible with António Damasio's neuroscientific theory of mind) to Luciano Floridi's *distributed morality*. But the main focus is on identity se the self, crowned within Francisco Varela's and Hum-

berto Maturana's *autopoiesis* moved even to the futurological dimension in projecting future trends in the 4th space.

The article by Christoph M. Abels, Daniel Hardegger on privacy and transparency in the 4th space may look like a very applied paper. In fact it does follow up on Abels' presentation and papers that are just coming out that focus on multifarious aspects of privacy and its problems in the virtual space. However, the article has also a second important aspect. It contains the most complete presentation to date, of the 4th space theory. While Daniel Hardegger came out with his interpretation of the 4th space in January 2021, those publications are often hard to locate and function as working projects. In this article we have a transparent, interesting and novel presentation of Hardegger's theory of the 4th space, strengthened y joint work and discussions within the 4ths space group. The article is worth reading for both the theoretical part and its practical application to the issues of privacy.

The following article, by Dustin Gray on virtual forms of surveillance and control is an interesting follow up on the work by Abels. While the topic is obviously relevant, some of the proposed solutions lead to philosophical a conundrum—where limits on control are seen as necessarily coming from our functioning in virtual space, or even in a large society.

## 5. EPISTEMOLOGY AND COMPUTERS

The name of Part 3—*Epistemology and Computers*—is a broad heading under which we placed somewhat more traditional articles on philosophical problems informed by computer science, especially AI. Magnus Johnsson opens this part with his article on perception, imagery, memory and consciousness. The author focuses on BICA approach; namely similarity of cognitive architectures between AI and animal/human brains. Johansson argues that some of the principles he puts forth are relevant for phenomenal consciousness of machines. He also develops such epistemic issues in AI, as memory, consciousness and imagination.

Rafał Maciąg writes about *knowledge as a phenomenon in the area of digital technologies, in particular artificial intelligence*. What seems like a standard article in epistemology opens up to the epistemologies generated by AI and related fields. This is followed by an article by Pavel N. Baryshnikov on computationalism in philosophy of mind. It uses anti-computationalist arguments to tackle the semantic problems, especially the lack of semantic properties, in in the computationalist theory of mind. This is relevant in various areas of AI.

The paper by Robin Hill is built on an interesting observation. There is an *artificialist fallacy*, defined as "causal justification of the *influence* of a technology, particularly artificial intelligence, by appeal to the *existence*

of the technology." This is akin to well-known naturalist fallacy. This lucid article also tackles the issue of value judgments in the artificialist fallacy.

Simon X. Duan, in his article on the "Platonic computer" tackles another conundrum—the inverse hard problem of consciousness. Idealism holds that consciousness is the fundamental nature of reality, thus, "matter is a derivative of consciousness." Thus, it becomes impossible to justify existence of the material world, which is the "inverse hard problem of consciousness" coined by Max Velmans. The paper is based on the concept of meta-computing and meta-consciousness as essential in generating "abstract entities" as well as "physical and nonphysical realities."

Last but not least, we have an article on dual-process approach to the problem of AI agency perception, by Marcin Rabiza. The author focuses on the two kinds of agency: 1. automatic, routine, often unconscious; 2. slower, controlled, more conscious. This is applied to AI.

This issue is maximally devoted to philosophy in computing, especially in AI, or sometimes philosophy in AI—not so much to philosophy *about* AI. This differentiates it from many other philosophical publications on AI, especially those from the 20th century. Philosophy seems much needed in theoretical AI, and many forms of cognitive science—while those disciplines open new conceptual avenues for philosophical thinking informed by developments, especially in artificial consciousness. This fruitful phase is just at its beginnings, as long as philosophers do not pontificate based on the old good theories and stay informed in the general trends, while engineers and scientists treat philosophy as an opening field for brainstorming and a design opportunity for the new views on reality.

*Piotr (Peter) Bołtuć*
Guest-editor
University of Illinois at Springfield, USA
The Warsaw School of Economics, Poland

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Mark Burgin, Rao Mikkilineni

# SEVEN LAYERS OF COMPUTATION: METHODOLOGICAL ANALYSIS AND MATHEMATICAL MODELING

doi: 10.37240/FiN.2022.10.zs.1

***ABSTRACT***

We live in an *information society* where the usage, creation, distribution, manipulation, and integration of information is a significant activity. Computations allow us to process information from various sources in various forms and use the derived knowledge in improving efficiency and resilience in our interactions with each other and with our environment. The general theory of information tells us that information to knowledge is as energy is to matter. Energy has the potential to create or modify material structures and information has the potential to create or modify knowledge structures. In this paper, we analyze computations as a vital technological phenomenon of contemporary society which allows us to process and use information. This analysis allows building classifications of computations based on their characteristics and explication of new types of computations. As a result, we extend the existing typologies of computations by delineating novel forms of information representations. While the traditional approach deals only with two dimensions of computation—symbolic and sub-symbolic, here we describe additional dimensions, namely, super-symbolic computation, hybrid computation, fused computation, blended computation, and symbiotic computation.

**Keywords:** symbol; structure; system; computation; process; symbolic; sub-symbolic; super-symbolic; superstructure; structural machine.

## 1. INTRODUCTION

The organization of computations in general and the power of operations of the utilized information processing devices have an indispensable impact on the efficiency of computations and the goals that this computation can achieve.

As people are accustomed to computing with symbols, the beginning of information processing technology started with the development of compu-

ting devices, which operated with symbols. This trend has been prevailing for quite a while. Therefore, all contemporary computers process symbolic information while digitalization expands to a variety of areas including computers, calculators, tablets, cell phones, TV sets, and servers to mention but a few. Similarly, the development of mathematical models of computation and algorithms started with symbolic systems such as Turing machines or partial recursive functions.

As a result, the symbolic computation was used for modeling the mind and its higher functions intelligence, and cognition, while experts in artificial intelligence formulated the Physical Symbol System Hypothesis (Newell, Simon, 1991), which stated: "A physical symbol system has the necessary and sufficient means for general intelligent action."

However, some researchers, for example, philosopher John Searle, criticized this hypothesis, while the development of information processing theory and technology brought forth another approach to computation and modeling higher brain functions (Searle. 1980). It was connectionism, which was later extended to associationism. According to connectionism, the functioning of the brain is based not on manipulation with symbols but on interactions of the highly connected network of neurons. At the same time, the model of artificial neural networks emerged as an alternative to symbolic information processing. At first, researchers did not regard the functioning of neural networks as computation but later the situation changed and it was assumed that it is subsymbolic or connectionist computation.

Adherents of the connectionist approach to computation in general and to AI in particular maintain that the level of symbolic information processing is too high for many problems and to elaborate an adequate model of the mind and build effective AI systems, it is necessary to utilize subsymbolic computation instead of designing programs that work with symbols.

Here we argue that it is crucial to organize computations not only on two levels—symbolic and subsymbolic but go higher to perform super-symbolic computations and combining all these forms to achieve a superior stage of performance and high level of intelligence. Thus, the goal of this paper is a methodological and philosophical analysis of different pure and combined or aggregated forms of computation.

This paper has the following structure. In Section 2, the concepts of s*ymbol* and *structure* are defined and analyzed. In Section 3, we discuss subsymbolic computations. In Section 4, we reflect on symbolic computations. In Section 5, we determine and explore super-symbolic computations. In Section 6, we describe tools created for operation with structures. Aggregated types of computation are elucidated in Section 7. In Conclusion, we discuss the results of this paper suggesting new directions for research in the theory and practice of algorithms and computation.

## 2. SYMBOLS AND STRUCTURES

Exploring of the utilization of the term "symbol" in contemporary society, it is possible to find that there are three main interpretations of the word "symbol:"

(1) symbol as a physical object with some meaning,

(2) symbol as a synonym of the concept *sign* being treated as a conceptual structure,

(3) symbol as a conceptual (theoretical) structure and a particular case of signs.

We will call a symbol by the name "*material symbol*" when we have in mind the first interpretation, by the name *conceptual sign* when we bear in mind the second interpretation, and *conceptual symbol* when we take into consideration the third interpretation. In computation, conceptual symbols are represented by material symbols.

Examples of material symbols are printed, written, and displayed on the screen letters, words, digits, and traffic signs.

According to David J. Chalmers, a symbol is an atomic entity, designating some object or concept, which can be manipulated explicitly by a physical symbol system, leading to intelligent behavior (Chalmers, 1992). Symbolic AI deals with the class of programs that perform computations directly upon such symbols.

An example of the case when the terms *symbol* and *sign* are treated as synonyms is the usage of the expressions *symbolic system* and *sign system* although the first one is used much more often.

The third meaning of the word *symbol* as a theoretical or philosophical structure is studied in semiotics as the science of signs. The name *semiotics* comes from ancient Greece where it was assumed that signs exist in nature while symbols function in society. Later Augustine of Hippo (354–430) determined *sign* as a general concept and *symbol* as its particular case in his study of signs and symbols (cf., (Deely, 2009)). An important contribution to this area was the book of John Poinsot (1589–1644) who was also called John of St. Thomas (Poinsot, 1632). The next imperative contribution to semiotics was done by Charles Sanders Peirce (1839–1914) in the form of a general theory of signs (Peirce, 1931–1935; Alp, 2010; Burgin, 2012; 2016; Burgin and Schumann, 2006; Goodman, 1968). Similar to Augustine of Hippo, Peirce treated *sign* as the general term while *symbol* as the convention-based sign constructing the following triadic model of a sign, *Balanced Sign Triad* (cf. Figure 1), where a sign is understood as a relation consisting of three elements: vehicle, object of the sign and meaning.

The Existential Triad of the world (Burgin, 2012) imposes the existence of three *substantial types* of signs and symbols: *material*, *mental,* and *con-*

*ceptual* signs/symbols, which belong to the physical world, mental word, and the structural world, correspondingly.

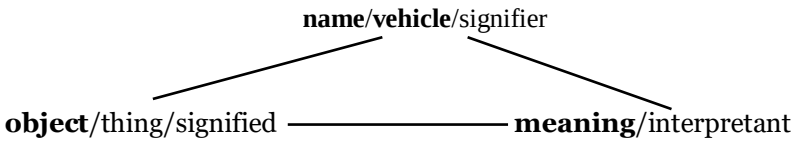



**Figure 1**. The Balanced Sign Triad or Sign Triangle of Peirce

Usually, when people speak or write about symbols, they mean material symbols. Note that material symbols are not only as individual aggregates of points, such as *a* or 3, written or printed on paper or displayed on the screen. Electrical charges, stones or pebbles can be also material symbols of numbers.

The Balanced Sign Triad of Peirce agreeably correlates with the Existential Triad of the world, which is formed of three basic components: the Physical World, the Mental World, and the World of Structures (Burgin, 2012). In Peirce's triad, the *name* corresponds to the World of Structures as a syntactic system, the *object*/thing corresponds to the Physical World, and the *meaning*/interpretant corresponds to the Mental World as a semantic system. At the same time, the *object* can be non-material and thus, beyond the Physical World. Nevertheless, the object is always closer to the Physical World. On the one hand, this implies that the Balanced Sign Triad of Peirce is homomorphic to the Existential Triad of the world, while on the other hand, it demonstrates fractality of the Existential Triad of the world, which is repeated in a diversity of other natural and artificial systems.

We remind that a *fractal* is a complex system displaying self-similarity across different scales (cf., for example, (Mandelbrot, 1983; Edgar, 2008)). In other words, *fractality* means that the structure of the whole is repeated/reflected in the structure of its parts on many levels (cf., for example, (Coleman and Pietronero, 1992; Calcagni, 2010)). It is possible to find formalized mathematical definition of fractals in (Lapidus, et al, 2017).

According to Peirce, there are three *relational types* of signs: *icon*, *index*, and *symbol*. Thus, as a particular case of signs, a conceptual symbol has the sign triad of Peirce is its structure.

**Definition 2.1.** An *icon* is an image of the object it signifies.

Photographs at the level of direct resemblance or likeness are prototypical examples of icons. Computer icons helped popularize the word being, as well as the pictographs such as those used on "pedestrian crossing" signs, typical examples of icons. There is no real connection between an object and its icon other than the likeness, so the mind itself is required to see the similarity and associate the two. A characteristic of the icon is that by observing

it, we can derive information about its object. The more simplified the image, the less it is possible to learn. No other kind of signs gives that kind of information.

Peirce further divides icons into three kinds:

— *images* have the simplest quality, the similarity of aspect, while portraits, photographs, and computer icons are images.
— *diagrams* represent relationships of parts rather than tangible features, while block schemes, flowcharts, and algebraic formulae are diagrams.
— *metaphors* possess a similarity of character, representing an object by using parallelism in some other object being widely used in poetry and language.

One more type of signs *index* has a causal and/or sequential relationship to its object. A key to understanding indices (or indexes) is the verb "indicate," of which "index" is substantive. For instance, directly perceivable events that can act as a reference to events that are not directly perceivable, or in other words, something visible that indicates something out of sight, are indices. You may not see a fire, but you do see the smoke and that indicates to you that a fire is burning and the smoke is its index. Such words as *this*, *that*, *these*, and *those* are also indices. The nature of the index can be unrelated to that of the signified, but the connection with it is logical and organic, e.g., the two elements are inseparable, and there is little or no participation of the mind to see this connection. Indices are generally non-deliberate, although written or printed arrows are just one example of deliberate ones.

A *symbol* represents something in a completely arbitrary relationship with its object. The connection between the signifier/name and signified/object depends entirely on the observer, or more exactly, what the observer was taught. Symbols are subjective. Their relation to the signified object is dictated either by social and cultural conventions or by habit. Words are the best example of symbols. Whether as a group of sounds or a group of characters, they are only linked to their signified because people decide they are  and because the connection is neither physical nor logical, words change the meaning or objects change names as time goes by. Here it all happens in mentality and depends on it.

Note that contrary to Peirce, the French linguist Ferdinand de Saussure (1857–1913) understood the concept *sign* as a subcategory of the concept *symbol*. This relation is represented by the *Dyadic Sign Triad* of Saussure and is presented in Figure 2 (Saussure, 1916).

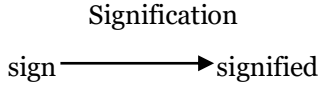
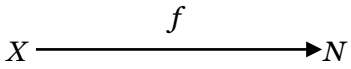



**Figure 2.** The *Dyadic Sign Triad* of Saussure

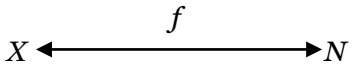
Note that this triad is a kind of the fundamental triad (Burgin, 2011).

Indeed, we have the following definition.

**Definition 2.2.**(a) A *basic named set*, also called a *basic fundamental triad*, is a triad $\mathbf{X} = (X, f, N)$ with the following visual (graphic) representation:




(b) A *bidirectional named set*, also called a *bidirectional fundamental triad*, is a triad $\mathbf{X} = (X, f, Z)$ with the following visual (graphic) representation:




The theory of named sets provides unified foundation of mathematics encompassing set theory, logic, category theory and homotopy type theory as its subtheories (Burgin, 2004). Moreover, it is proved that all mathematical structures, e.g., functions, relations, graphs, categories, functors, operators, and topological spaces, are either named sets or systems of named sets (Burgin, 2011).

Returning to the concept of sign, we see that in contrast to de Saussure and Peirce, Morris defines *sign* in a dynamic way relative to some interpreter. He writes that *S* is a sign (the sign name) of an object or objects *D* for an interpreter *I* to the degree that *I* takes the account of *D* in virtue of the presence of *S* (Morris, 1938). Thus, the object *S* becomes a sign only if somebody (an interpreter) interprets *S* as a sign (the sign name). This gives us the following diagram.

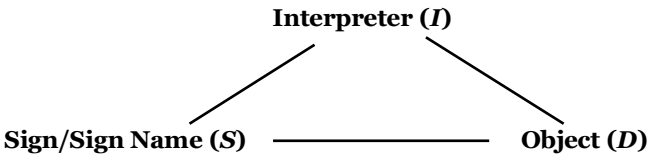



**Figure 3**. The *Dynamic Sign Triad* of Morris

According to Morris, the sign name is what supports the triadic relation of the sign with other signs, with designated objects and with the subjects using the sign. These relations are represented by the corresponding fields of semiotics.

The Dynamic Sign Triad of Morris also correlates with the Existential Triad of the world. In Morris' triad, the *name* corresponds to the World of Structures as a syntactic system, the *object* can be associated with the Physi-

cal World, and the *Interpreter* corresponds to the Mental World as a system mentality that comprehends *S* as a sign (symbol). At the same time, the *object* can be non-material and thus, beyond the Physical World. Nevertheless, the object is always closer to the Physical World. On the one hand, this implies that the Peircean triad is homomorphic to the Existential Triad of the world, while on the other hand, it demonstrates fractality of the Existential Triad of the world, which is repeated in a diversity of other natural and artificial systems.

Observing symbols and signs, we can see their inherent relation to concepts. Indeed, let us look at the Concept Triangle of Russell presented in Figure 4 (Russell, 1905). Then taking Denotation as an Object, we come from the Concept Triangle of Russell to the Balanced Sign Triad (Sign Triangle) of Peirce while interpreting Object as Denotation, we come from the Sign Triangle of Peirce to the Concept Triangle of Russell.

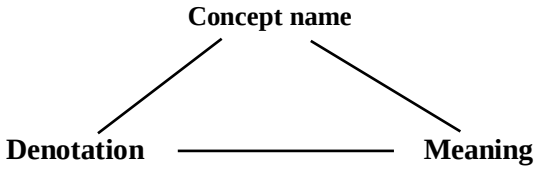



**Figure 4.** The *Concept Triangle* of Russell

The most advanced model of concepts—the Representational Triad—presented in Figure 5 is homomorphic to the Dyadic Sign Triad of Saussure (Burgin and Gorsky, 1991; Burgin, 2012). At the same time, both of them are particular cases of the fundamental triad.

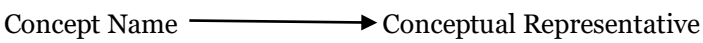



**Figure 5.** The Representational Triad of a Concept

Signs, symbols, and concepts are structures. Computations are performed by operation with structures. That is why we present the exact definition of a structure.

Traditionally it is defined as follows (cf., for example, (Robinson, 1963; Grossmann, 1990; Tegmark, 2008)):

**Definition 2.3.** A *structure* is (a representation of) a complex entity that consists of parts in relations to each other.

However, it was demonstrated that this definition is essentially incomplete. A more comprehensive formal definition was suggested by Bourbaki in the case of mathematical structures (Bourbaki, 1957).

It is necessary to remark that Bourbaki (1957; 1960) also elaborated a formal definition of a general mathematical structure as a very abstract concept. Here we provide a short description of the formal definition from (Bourbaki, 1960) omitting some formal details and expressions.

Bourbaki start their definition of a structure $\Sigma$ with a finite sequence of ordered pairs of whole numbers, calling this sequence an *echelon construction scheme*. Then taking such a scheme $S$ and $n$ terms $E_1, E_2, E_3, \dots, E_n$, which denote (name) sets, in a formal theory **T** that is stronger that a theory of sets, such as ZF, they build an *echelon construction* $E$ of the scheme $S$ on the sets $E_1, E_2, E_3, \dots, E_n$, which are taken as the building blocks of the inductive construction employed by Bourbaki. Each step of this construction consists either of taking the Cartesian product $(E \times F)$ of two sets obtained in the preceding steps or of taking the power set $2^D$ of the set $D$ obtained in the previous steps. This echelon construction of the scheme $S$ is a sequence of terms in the theory **T** built according to the scheme $S$. After building this construction, Bourbaki take (in the theory **T**) a formal representation of a group of mappings $f_i$: $E_i \rightarrow E_i'$ $(i = 1, 2, 3, \dots, m)$ and determine *canonical extensions* with the scheme $S$ of the mappings $f_i$ to a mapping of an echelon construction $E$ of the scheme $S$ on the sets $E_1, E_2, E_3, \dots, E_n$.

After this, Bourbaki characterize a *typification* $T$ of letters $x_1, x_2, x_3, \dots, x_n$ in **T**. Subsequently, they delineate the concept of a *transportable relation* with respect to $T$. Next Bourbaki define (1) a *species of the structure* $\Sigma$ in **T** as a text that is a combination of letters $x_1, x_2, x_3, \dots, x_n, s$, terms in **T**, (2) a typification $T\{x_1, x_2, x_3, \dots, x_n, s\}$ of the letters $x_1, x_2, x_3, \dots, x_n, s$ in **T** called the *typical characterization* of the species of the structure $\Sigma$, and (3) a relation $R\{x_1, x_2, x_3, \dots, x_n, s\}$ that is transportable with respect to $T$ and called the *axiom* of the species of the structure $\Sigma$.

To define a "species of structure" $\Sigma$, Bourbaki take:

(1) $n$ sets $E_1, E_2, \dots, E_n$, as "principal base sets."

(2) $m$ sets $A_1, A_2, \dots, A_m$, the "auxiliary base sets", and finally

(3) a specific echelon construction scheme $S(X_1, X_2, \dots, X_n, A_1, A_2, \dots, A_m)$.

All these auxiliary constructions and definitions allow Bourbaki to define the *structure of species* $\Sigma$, taking terms $E_1, E_2, E_3, \dots, E_n$ in the theory **T** as principal base sets. Namely, this construction leads to the following definition (Bourbaki, 1960).

**Definition 2.4.** A term $U$ in the theory **T** is called a *structure of species* $\Sigma$ if the relation

$$T\{E_1, E_2, E_3, \dots, E_n, U\} \& R\{E_1, E_2, E_3, \dots, E_n, U\}$$

is a theorem in **T**.

Covering several pages in the book (Bourbaki, 1960), the completely formalized formal definition of a structure in the sense of Bourbaki is essentially much more complex and much longer than the partially formal definition

given above. Besides, this definition is too abstract and complicated even for the majority of mathematicians, who prefer to use an informal notion of a mathematical structure or the definition where a structure is formalized as a set with relations in this set. As Corry (1996) writes, Bourbaki's concept of *structure* was, from a mathematical point of view, a superfluous undertaking. Even Bourbaki themselves did not use this formalized concept in their later books of the *Eléments* after they had introduced it in *Theory of Sets* (Bourbaki, 1960). However, being overcomplicated, this definition is still incomplete. For instance, this definition does not discern inner and outer structures.

The complete formal definition of a structure was developed in the general theory of structures including both formal and informal forms (Burgin, 2012). Here we give only an informal definition of set-theoretical structures. There are also mereological structures, which are defined and explored in (Burgin, 2012).

**Definition 2.5.** A *structure R* consists of elements/parts and connections/relations of the three categories:

— Connections/relations between (groups of) elements/parts
— Connections/relations between (groups of) elements/parts and (groups of) connections/relations
— Connections/relations between (groups of) connections/relations

Note that elements themselves can be and often are structures. This property is called nesting and is used in many processes to improve their efficiency (Burgin, 2020).

Structures, elements of which are also structures, are called *superstructures*.

According to the general theory of structures (Burgin, 2012), there are three *existential types* of structures:

1. Ideal structures
2. Abstract structures
3. Embedded structures

Embedded structures are structures of physical systems (things), such as tables, trees or cars, and of mental systems, such as thoughts or values. An interesting example of a structure embedded in the mentality of society is a moral space studied in (Boltuc, 2013).

Abstract structures exist in the mentality of people or groups of people and are characterized only by their properties.

Ideal structures dwell in the world of structures described for example in (Burgin, 2017).

This world is the scientific incarnation of the world of Plato Ideas (Forms). Indeed, for millennia, the enigma of the world of Ideas or Forms, which Plato suggested and advocated, has been challenging the most prominent thinkers of the humankind. The solution to this problem was found

only recently. Namely, an Idea/Form in the Plato's sense can be interpreted as a scientific object called a structure. The difference is that Plato Ideas/Forms do not have a rigorous inclusive definition while structures have an accurate definition in the general theory of structures (Burgin, 2012). Based on this definition, it was demonstrated that structures have the basic properties of Plato's Ideas. In addition, it was proved the existence the world of ideal structures (Burgin, 2017).

It is also important that it was possible to discover the most basic atomic structure in the world of structures. It is called *fundamental triad* or *named set* (Burgin, 2011). Its definition is given in Section 2. Any structure is either a fundamental triad (named set) or is built of some number of fundamental triads (named sets). It means that the discovery of fundamental triad (named set) actually accomplished the search at first of philosophers and later of physicists for the entity out of which everything in the world is built. In this sense, the theory of named sets is the theory of everything (Burgin, 2011).

Assessing the place of structures in the world and their roles, it was found in the general theory of structures (Burgin, 2012) that systems have five substantive types of structures: *inner*, *internal*, *intermediary*, *outer,* and *external* structures.

According to the general theory of structures, we have the following definitions of these types:

**Definition 2.6.** (a) An *internal structure TQ* of a system *R* contains only inner structural parts, components and elements, i.e., parts, components and elements of *R*, relations between these parts, components and elements, relations between these parts, components, elements and relations from *TQ* and relations between relations from *TQ*.

(b) An *inner structure IQ* of a system *R* is a substructure of an internal structure *TQ* of *R*, where *IQ* is obtained by exclusion of (1) the whole system *R* as a part, component or element of itself and (2) all relations that include *R*.

(c) An *external structure EQ* of a system *R* is an extension of the internal structure, in which other systems, their parts, components and elements are included, as well as relations between all these included parts, components and elements, relations between these parts, components, elements and relations from *EQ* and relations between relations from *EQ*.

(d) An *intermediate structure MQ* of a system *R* is a substructure of an external structure *EQ* of *R*, where *MQ* is obtained by exclusion of (1) the whole system *R* and other systems from *EQ*, as well as (2) all relations that include these systems.

(e) An *outer structure OQ* of a system *R* is an inner structure of a system *U* in which *R* is only one of the inner elements of the inner structure *IQ* of the system *U*.

It is also possible to classify structures by their elements. It results in three pure classes of structures:

— *Subsymbolic structures* have only elements that are not treated as symbols but as parts of symbols
— *Symbolic structures* have only elements that are treated as symbols
— *Super-symbolic structures* have elements that are assembled from symbols

Examples of subsymbolic structures are sets of pixels on the screen of a computer, tablet or TV set.

Symbolic structures are composed of symbols in a simple way, that is, these structures have low structural complexity. Symbols, words, texts as a linear composition of words, and sets are symbolic structures.

Letters and digits are paradigmatic examples of symbols while words and numerals are examples of symbolic structures.

Texts, hypertexts, and diagrams are examples of super-symbolic structures.

In addition, there are *mixed structures*, which have elements of both types:

— *Hybrid structures* have both elements that are comprehended as symbols and elements that are treated as parts of symbols
— *Fused structures* have both elements that are recognized as symbols and elements that are operated as super-symbolic structures, for example, as assemblages of symbols
— *Blended structures* have both elements that are identified with subsymbols and elements that form super-symbolic structures, for example, as assemblages of symbols
— *Symbiotic structures* have elements of all three pure types

According to the general theory of structures, there is a hierarchy of structures composed of different orders of structures (Burgin, 2017).

Let us consider the mathematical formalization of the two first levels of this hierarchy.

**Definition 2.7** (Burgin, 2012)**.** A *first-order structure* is a triad of the form

$$\boldsymbol{A} = (A, r, \boldsymbol{R})$$

In this expression, we have:

— the set $A$, which is called the *substance* of the structure $\boldsymbol{A}$ and consists of elements of the structure $\boldsymbol{A}$, which are called *structure elements* of the structure $\boldsymbol{A}$

— the set $\boldsymbol{R}$, which is called the *arrangement* of the structure $\boldsymbol{A}$ and consists of relations between elements from $A$ in the structure $\boldsymbol{A}$, which have the first order and are called *structure relations* of the structure $\boldsymbol{A}$

— the *incidence relation r*, which connects groups of elements from $A$ with the names of relations from $\boldsymbol{R}$

**Examples of structures of the first order**:

The order relation: $1 < 2 < 3 < 4 < 5$
A string: $a - b - c - d - e$
A word: $s - e - v - e - n$

**Definition 2.8** (Burgin, 2012)**.** A *second-order structure* is a triad of the form

$$\boldsymbol{A} = (A, r, \boldsymbol{R})$$

Here

— the set $A$, which is called the *substance* of the structure $\boldsymbol{A}$ and consists of elements of the structure $\boldsymbol{A}$, which are called *structure elements* of the structure $\boldsymbol{A}$

— the set $\boldsymbol{R}$, which is called the *arrangement* of the structure $\boldsymbol{A}$ and consists of relations in the structure $\boldsymbol{A}$, which are called *structure relations* of the structure $\boldsymbol{A}$

— $r$ is the incidence relation that connects groups of elements from $A$ and/or relations from $\boldsymbol{R}$ with names of relations from $\boldsymbol{R}$

— $\boldsymbol{R} = \boldsymbol{R}_1 \cup \boldsymbol{R}_2 \cup \boldsymbol{R}_3$

— $\boldsymbol{R}_1$ is the set of relations between the elements from the set $A$

— $\boldsymbol{R}_2$ is the set of relations in the set $\boldsymbol{R}_1$, i.e., elements from $\boldsymbol{R}_2$ are relations between relations from $\boldsymbol{R}_1$

— $\boldsymbol{R}_3$ is the set of relations between elements from $A$ and relations from $\boldsymbol{R}_1$

Relations from $\boldsymbol{R}_2$ and $\boldsymbol{R}_3$ are called relations of the second order in $\boldsymbol{A}$.

Second-order structures are used to represent data processed by structural machines of the second order.

Similarly, we determine relations and structures of higher orders.

**Examples**:

1. The strict order $<$ on numbers is a suborder of the non-strict order $\leq$ on numbers. This is a relation of the second order.

2. Addition and subtraction are ternary relations. Subtraction is inverse to addition. This is a relation of the second order.

3. A function is a binary relation. When one function is an extension of another function, it defines a relation of the second order.

4. 0 is neutral element with respect to addition. This is a relation of the second order.

5. Taking relations between people, when the relations between A and B are better than the relations between A and D, it defines a relation of the second order.

In information technology, supercomputers are computers that have essentially better characteristics of information processing in comparison with ordinary computers. Usually, the improved characteristic is the higher speed of computing. In a similar way, superstructures are structures that have essentially higher complexity.

Symbolic superstructures are composed from symbols and symbolic structures. Intricate hypertexts, operational schemas, multicomponent images, and structures of higher order are symbolic superstructures.

Now let us analyze how operating with different types of structures shape specific types of information processing in general and computation in particular.

## 3. SUBSYMBOLIC COMPUTATIONS IN NATURE AND ARTIFICIAL DEVICES

Although there are different definitions of subsymbolic computation, here we uphold the following definition.

**Definition 3.1.** *Subsymbolic computation* is computation in which elements of processed data are not interpreted as symbols or sets of symbols by the computing system.

This well correlates with the opinion that in a system performing subsymbolic computation, the objects of computation are more fine-grained than the objects of semantic interpretation (Chalmers, 1992).

For instance, it is often assumed that the tokens manipulated by neural networks performing primitive operations are subsymbolic as they are located at a level lower than that of the symbols (Rumelhart, McClelland, 1986; Smolensky, 1988). Examples of such tokens are the activation values of neurons. In many cases, in a system performing subsymbolic computation, the computational level lies beneath the representational level (Chalmers, 1992).

Starting from the last quarter of the 19th century, there was an assumption in computer science, artificial intelligence and cognitive sciences that neural networks, which represent the connectionist model, perform subsymbolic computations. For instance, Smolensky introduced the following Subsymbolic Hypothesis as "the cornerstone of the subsymbolic paradigm":

> "The intuitive processor is a subconceptual connectionist dynamical system that does not admit a complete, formal, and precise conceptual level description." (Smolensky, 1988)

A paradigmatic example of subsymbolic computations is provided by artificial neural network, which lately became extremely popular due to their ability to perform deep learning.

Analog computing is another important form of subsymbolic computations.

In comparison with symbolic computation, subsymbolic computation has the following advantages (Kwasny, Faisal, 1992):

— It is more robust in noisy conditions
— Provides better performance for analog data
— It demands less knowledge upfront
— It is easier for scaling up
— It better adapts to Big Data
— It is better for perceptual problems
— It is more useful for building models in neuroscience

Indeed, now the prevailing opinion of neuroscientists is that intuitive mental processes that internal functioning of the brain does not utilize a symbolic description but require subsymbolic descriptions inherent for connectionist architecture. As a result, the subsymbolic paradigm provides better means for modeling the capabilities of the brain, which also implies reduction of mental to neural computation.

## 4. SYMBOLIC COMPUTATIONS AS THE BASIC FORM OF ALGORITHMIC INFORMATION PROCESSING

Now we come to symbolic computations. In contrast to subsymbolic computation, in a system performing subsymbolic computation, the objects of computation are also objects of semantic interpretation and very often the computational level coincides with the representational level (Chalmers, 1992).

**Definition 4.1.** *Symbolic computation* is computation in which elements of processed data are interpreted as symbols by the computing system and computation is performed by the individual transformation of these symbols.

Turing machines are the paragon of automata performing symbolic computations. Indeed, on each step of its computation a Turing machine, observes a symbol in a cell of its tape and eventually changes this symbol to another symbol before moving to another cell.

In comparison with subsymbolic computation, symbolic computation has the following advantages (Kwasny, Faisal, 1992):

— In it, introspection more useful for coding
— It is easier to debug
— It is easier to understand and explain
— It is easier to control
— It is more efficient for solving abstract problems
— It is better adapted for explaining people's thinking

At the same time, the symbolic/subsymbolic distinction does not imply the architectural dissimilarity (Chalmers, 2018). Indeed, on the one hand, Turing machines can and contemporary digital computers do simulate neural networks, which are paradigm examples of subsymbolic computation. On the other hand, neural-network can precisely model Turing machines (Siegelman, 1999).

Connectionist paradigm also does not contradict the possibility to imply it using symbolic computation. Indeed, cellular automata have the connectionist architecture but each cell of these automata performs symbolic computation (cf., for example, (Burgin, 2005)).

Now we can describe the new type of computation.

## 5. SUPER-SYMBOLIC COMPUTATIONS AS A NEW DIMENSION OF INFORMATION PROCESSING

Analysis of real-life computations show that computers and other advanced information processing systems, such as the brain, operate not only with symbolic and subsymbolic data but also with essentially more advanced structures.

**Definition 5.1.** Computation is *super-symbolic* when symbolic structures of higher orders and superstructures are transformed as holistic objects.

This contrasts symbolic computations where symbolic structures are transformed by operating with separate symbols.

Super-symbolic (transcendent) computation is a model of functioning of the right hemisphere of the brain. Indeed, processing images of material systems by transformations of holistic shapes is an example of super-symbolic computation.

One more important example of super-symbolic computation is operation with schemas (Burgin, Mikkilineni, 2021). These processes are very important for the functioning of the mind because the framework of schema theory can provide a better bridge from human psychology to brain theory than that offered by the symbol systems (Arbib, 2021).

The advantage of the super-symbolic (transcendent) computing is its ability to operate big formal and informal systems of data and knowledge with high efficiency. That is why the implementation of super-symbolic computing is the way to the solution of the problem of big data and information overflow.

## 6. AMALGAMATED LEVELS OF COMPUTATION

The combinations of pure types give mixed types of information processing. The first step in this direction gives us *hybrid computation*, which comprises both symbolic and subsymbolic computations being a two-fold type of computations (Burgin, Dodig-Crnkovic, 2015). Hybrid computation allows combining advantages of both symbolic and subsymbolic computations.

Researchers found that individual neurons can perform symbolic computations (Cepelewicz, 2020; Gidon, et al, 2020). For instance, it was discovered that individual dendritic compartments can also perform a particular computation—"exclusive OR"—that mathematical theorists had previously categorized as unsolvable by single-neuron systems.

Moreover, some psychologists assume that the roots of arithmetic reside in single neurons (Dehaene, 2002). It means that neural networks in the brain perform both symbolic and subsymbolic computations, i.e., they operate is on the level of hybrid computation.

Conventional models of computation perform either symbolic computation, e.g., finite automata, Turing machines, inductive Turing machines or Random Access Machines (RAM), or subsymbolic computation, e.g., neural networks or cellular automata. New models, such as neural Turing machines (Graves, et al, 2014; Collier, Beel, 2018) or structural machines with symbolic and subsymbolic processors, carry out hybrid computation.

A neural Turing machine is a recurrent neural network with a network controller connected to external memory resources. As a result, it combines subsymbolic computation of neural networks with symbolic computation of Turing machines.

Super-symbolic (intuitive) computation adds one more dimension to the general schema of computational processes. This allows merging this type with already known types brings us to the system of *three twofold types* of computation:

— *hybrid computation* combines symbolic and subsymbolic computation

— *blended computation* combines subsymbolic and super-symbolic computation

— *fused computation* combines symbolic and super-symbolic computation

While it is easy to understand how information processing systems, such as computers or the brain, can perform fused computations, realization of blended computation looks more intriguing. One way to do this is simply to utilize two types of processors in the computing system—processors of one type work with subsymbolic data whereas processors of the other type oper-

ate super-symbolic data. Another mode of blended computation has two stages. At the first stage, the computing system processes subsymbolic input data developing one or several super-structures. These super-structures are handled at the second stage of the computational process. For instance, a neural network can aggregate or find an operational schema and then this schema is used and transformed, for example, improved, by an appropriate assembly of neural networks.

Synthesizing super-symbolic computation with symbolic (rational) computation and subsymbolic (intuitive) computation in one model, we come to *symbiotic computation*. Structural machines with flexible types of processors can accomplish symbiotic computation. Symbiotic computation allows combining advantages of all three pure types of computation representing the entire type of computations.

Thus, there is also one entire type of information processing:

— *symbiotic computation* combines all three pure types of information processing.

It is possible to consider symbiotic computation as the highest level of computation as it comprises all other types of computation.

## 7. OPERATING WITH STRUCTURES AND SCHEMAS

The identification of the new types of computation needs machines that would be able to perform such computations. Structural machines provide means for all types of computation including symbiotic computation when the machines possess processors of different types (Burgin, Adamatzky, 2017; Burgin, 2020). Let us describe these powerful models of computation.

A structural machine M works with structures of a given type and has three components:

1. The *control device* $C_M$ regulates the state of the machine $M$
2. The (entire) *processor* $P_M$ performs transformation of the processed structures and its actions (operations) depend on the state of the machine $M$ and the state of the processed structures. The entire processor consists of one or several unit processors. When a structural machine is considered only as a theoretical model, it is possible that the entire processor contains infinitely many unit processors.
3. The *functional space* $Sp_M$ consists of three components:
   - The *input space* $In_M$, which contains the input structure(s).
   - The *output space* $Out_M$, which contains the output structure(s).
   - The *processing space* $PS_M$, in which the input structure(s) is transformed into the output structure(s), which form the results of computation of a structural machine.

Unit processors can move in the processing space performing operations with structures in their neighborhoods according to the rules of their structural machine. Unit processors can function in the centralized mode when they are regulated by the common centralized control device. When the structural machine has the distributed control device, which consists of several unit control devices, the unit processors of this machine can function in two modes: clusterized and totally distributed modes. In the clusterized mode, all unit processors of the structural machine are divided into several groups (clusters) and each group works with its own control device. In a totally distributed mode, each unit processor has its individual control device. This architecture of the structural machine allows considerable flexibility and adaptivity.

Unit processors of one structural machine can be of different types and categories. For instance, it is possible that one unit processor is a Turing machine, another unit processor is a neural network, while the third one is a cellular automaton and one more unit processor is an inductive Turing machine.

It is natural to assume that all structures—the input structures, the output structures and all processed structures—have the same type.

The computation of a structural machine $M$ determines the *trajectory of computation*, which is a tree in general case and a sequence when the computation is deterministic and is performed by a single processor unit.

## 8. CONCLUSION

Information, and the computing structures that process it, play a critical role in how we, as humans, perceive the structural reality that surrounds us and how we interact with it. The Existential Triad of the world derived from the general theory of information describes the three worlds that interact with each other (Burgin, 2012). First, we have the material world, where structures exist and obey the laws of conversion of energy and matter. Biological systems have through evolution, and natural selection developed information processing structures that receive information about other material structures through various senses they have developed using their physical structures.

In addition, there is a mental world encompassing mental structures. Mental structures allow living systems to create and use their "vital potentialities and life processes." According to the general theory of information, knowledge derived from information can be represented in the form of ideal structures consisting of an atomic structure called the "fundamental triad." The fundamental triad (also known as a name set) consists of entities, relationships, and behaviors caused by actions and events that change the state

of the entities performing information processing in general and computation in particular.

In this paper, we have discussed two types of computational structures namely, symbolic and subsymbolic computation, which are ubiquitous in the current state of the art in information technologies. Symbolic computing deals with the evolution of structures made up of symbols and subsymbolic computing deals with elements of processed data that are not interpreted as symbols or sets of symbols by the computing system. Deep learning using a neural network model is an example.

In addition, we have also analyzed super-symbolic computation where structures as holistic objects are processed in contrast to symbolic computation where sequences of symbols are processed. This approach has many advantages (Burgin, Mikkilineni 2021) going beyond current symbolic and subsymbolic computational methods used in information technologies. The highest type—symbiotic computation allows us to use symbolic, subsymbolic, and super-symbolic computations. An application of symbiotic computing using symbolic, subsymbolic, and super-symbolic computing is discussed in Burgin-Mikkilineni Thesis (Burgin, Mikkilineni, 2021; Mikkilineni, 2022). The other three forms of computing are fused computation, blended computation, and hybrid computation, which are also observed in nature.

In this paper, we have presented a new and comprehensive picture of information structures, information processes (computations), and the associated tools derived from the general theory of information. We hope that this will guide us to not only understand how we as humans process and use information, but also will allow us to build a new class of digital automata that mimic how people process information.

**Acknowledgments:** The authors would like to express their gratitude to Piotr Boltuc for useful remarks.

## REFERENCES

M. A. Arbib, *Schemas Versus Symbols: A Vision from the 90s*, Journal of Knowledge Structures and Systems, 2021, 2 (1), pp. 68–74.

K. O. Alp, *A Comparison of Sign and Symbol (Their Contents and Boundaries)*, *Semiotica*, 2010, 182(1–4), pp. 1–13.

D. S. Blank, Meeden, L.A., Marshall, J., *Exploring the Symbolic/Subsymbolic Continuum: A Case Study of RAAM*, in: *The Symbolic and Connectionist Paradigms: Closing the Gap*, Psychology Press, 1992.

P. Boltuc, *Non-homogenous Moral Space: From Bentham to Sen*, Analiza i Egzystencja, 2013, 24, pp. 43–59.

N. Bourbaki, *Structures*, Hermann, Paris, France 1957.

_____ , *Theorie Des Ensembles,* Hermann, Paris 1960.

M. Burgin, *Unified Foundations of Mathematics*, Preprint Mathematics LO/0403186, 2004, 39 p.; electronic edition: http://arXiv.org.

_____ , *Superrecursive Algorithms*, Springer, New York 2005.
_____ , *Theory of Named Sets*, Mathematics Research Developments, Nova Science, New York, 2011
_____ , *Structural Reality*, Nova Science Publishers, New York 2012.
_____ , *Theory of Knowledge: Structures and Processes*, World Scientific, New York–London–Singapore 2016.
_____ , *Ideas of Plato in the Context of Contemporary Science and Mathematics*, Athens Journal of Humanities and Arts, 2017, 4 (3), pp. 161–182.
_____ , *Information Processing by Structural Machines*, in: Theoretical Information Studies: Information in the World, World Scientific, New York–London–Singapore 2020, pp. 323–371.
_____ , *Elements of the Theory of Nested Named Sets*, Theory and Applications of Mathematics & Computer Science, 2020a, 10 (2), pp. 46–70.
M. Burgin, A. Adamatzky, *Structural Machines and Slime Mold Computation*, International Journal of General Systems, 2017, 45 (3), pp. 201–224.
M. Burgin, G. Dodig-Crnkovic, *A Taxonomy of Computation and Information Architecture*, in: Proceedings of the 2015 European Conference on Software Architecture Workshops, Dubrovnik–Cavtat, Croatia, September 7–11, 2015, ACM, pp. 71–78.
M. Burgin, D. Gorsky, *Towards the Construction of General Theory of Concept*, in: The Opened Curtain, Oulder–San Francisco–Oxford, 1991, pp. 167–195.
M. Burgin, R. Mikkilineni, *From Data Processing to Knowledge Processing: Working with Operational Schemas by Autopoietic Machines*, Big Data Cogn. Comput., 2021, 5 (13); https://doi.org/10.3390/bdcc5010013
_____ , *On the Autopoietic and Cognitive Behavior*, EasyChair Preprint no. 6261, version 2, 2021; https://easychair.org/publications/preprint/tkjk
M. Burgin, R. Mikkilineni, V. Phalke, *Autopoietic Computing Systems and Triadic Automata: The Theory and Practice*, Advances in Computer and Communications, 2020, 1 (1), pp. 16–35.
M. Burgin, J. H. Schumann, *Three Levels of the Symbolosphere*, *Semiotica*, 2006, 160(1–4), pp. 185–202.
G. Calcagni, *Fractal Universe and Quantum Gravity*, *Phys. Rev. Lett*., 2010, 104, 251301.
D. J. Chalmers, *Subsymbolic Computation and the Chinese Room*, in: *The Symbolic and Connectionist Paradigms: Closing the Gap*, Lawrence Erlbaum, 1992, pp. 25–48.
J. Cepelewicz, *Hidden Computational Power Found in the Arms of Neurons*, Quanta Magazine, 2020.
P. Coleman, L. Pietronero, *The Fractal Structure of the Universe*, *Phys. Rep*., 1992, 213, 311.
M. Collier, J. Beel, *Implementing Neural Turing Machines*, Artificial Neural Networks and Machine Learning – ICANN 2018, 2018, pp. 94–104.
Corry, L. *Modern Algebra and the Rise Mathematical Structures*, Birkhäuser, Basel–Boston–Berlin 1996.
J. Deely, *Augustine & Poinsot: The Protosemiotic Development*, University of Scranton Press, Scranton 2009.
S. Dehaene, Single-neuron Arithmetic, *Science*, 2002, 297, pp. 1652–1653.
G. Dodig-Crnkovic, *Cognition as Embodied Morphological Computation, in Philosophy and Theory of Artificial Intelligence*, Springer, New York 2017, pp. 19–23.
G. Edgar, *Measure, Topology, and Fractal Geometry*, Springer, New York 2008.
N. Goodman, *Languages of Art*: *An Approach to a Theory of Symbols*, Bobbs-Merrill, Indianapolis 1968.
A. Graves, G. Wayne, I. Danihelka, *Neural Turing Machines*, arXiv:1410.5401, 2014.
R. Grossmann, *The Fourth Way*: *A Theory of Knowledge*, Indiana University Press, Bloomington, Indianapolis 1990.
S. C. Kwasny, K. A. Faisal, *The Symbolic and Connectionist Paradigms: Closing the Gap*, Lawrence Erlbaum 1992.
M. L. Lapidus, G. Radunović, D. Žubrinić, *Fractal Zeta Functions and Fractal Drums: Higher-Dimensional Theory of Complex Dimensions*, Springer Monographs in Mathematics, Springer, New York 2017.
B. Mandelbrot, *The Fractal Geometry of Nature*, Macmillan, New York 1983.

R. Mikkilineni, *A New Class of Autopoietic and Cognitive Machines*, Information, 2022, 13, 24; https://doi.org/10.3390/info13010024
C. W. Morris, *Foundation of the Theory of Signs*, in: International Encyclopedia of Unified Science, 1938, 1 (2).
A. Newell, H. A. Simon, *Computer Science as Empirical Inquiry*, Communications of the Association for Computing Machinery, 1991, 19, pp. 113–126.
J. Poinsot, *Tractatus de Signis*, Alcala de Henares, 1632.
A. Robinson, *Introduction to Model Theory and Metamathematics of Algebra*, North-Holland, Amsterdam–New York 1963.
D. Rumelhart, J. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. l. Foundations, MIT Press, Cambridge, MA 1986.
B. Russell, *On Denoting*, Mind, 1905, 14, pp. 479–493.
F. de Saussure, *Nature of the Linguistics Sign*, in: Cours de linguistique générale, McGraw Hill Education, 1916a.
J. R. Searle, *Minds, Brains and Programs,* Behavioral and Brain Sciences, 1980, 3, pp. 417–424.
H. T. Siegelman, *Neural Networks and Analog Computation*: *Beyond the Turing Limit,* Birkhauser, Berlin 1999.
P. Smolensky, *On the Proper Treatment of Connectionism*, Behavioral and Brain Sciences, 1988, 11 (1), pp. 1–23.
_____ , *Subsymbolic Computation Theory for the Human Intuitive Processor*, in: Proc. *Computability in Europe Conference* 2012. *How the World Computes*, LNCS 7318, 2012, pp. 676–686.
M. Tegmark, *The Mathematical Universe*, Foundations of Physics, 2008, 38 (2), pp. 101–150.

ABOUT THE AUTHORS —

Mark Burgin — UCLA, Los Angeles, California, USA
Email: mburgin@math.ucla.edu

Rao Mikkilineni — Golden Gate University, San Francisco, USA
Email: rmikkilinni@ggu.edu

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Piotr (Peter) Bołtuć

# NON-REDUCTIVE PHYSICALISM FOR AGI

doi: 10.37240/FiN.2022.10.zs.2

*To the memory of Gilbert Harman*[1]

***ABSTRACT***

Creature consciousness provides a physicalist account of the first-person awareness (*contra* Rosenthal). I argue that non-reductive consciousness is not about phenomenal qualia (Nagel's *what it is like to feel like something else*); it is about the stream of awareness that makes any objects of perception epistemically available and ontologically present. This kind of consciousness is central, internally to one's awareness. Externally, the feel about one's significant other's that "there is someone home" is quite important too. This is not substance dualism since creature consciousness and functional consciousness are both at different generality levels of physicalism. Surprisingly, pre-Hegel philosophy of pure subject is more fitting with the current engineering approach than analytical phenomenalism. The complementary view of subject- and object-related perspectives, may come from Fichte's *Wissenschaftslehre*; but here it is placed, securely within the physicalist paradigm. It is essential to the Engineering Thesis in Machine Consciousness, which helps us understand under what general conditions a machine would be first-person conscious, but when it is merely functionally conscious.

**Keywords**: Machine consciousness, non-reductive physicalism, non-reductive machine consciousness, creature consciousness, non-reductive consciousness; complementary philosophy, *Wissenschaftslehre,* two-tier physicalism.

---

[1] This article is dedicated to the memory of Gilbert Harman, who has shown me at least three things: how to do analytical philosophy without outdoing on analyticity; how to reach beyond the boundaries of philosophy and gently disrespect those boundaries; how to treat a student with true respect. When I have asked Gil about the latter, he just said: Those are my colleagues who were born a bit later than I was. Several of them are to become stronger philosophers than I am. Why would I not treat them as such? *Oh well, Gill has also shown me how not only to love, but also truly respect, one's daughters; me and my five years old are grateful for this.*

## 1. INTRODUCTION

In Part I we focus on philosophical issues that relate to physicalism. The goal is to explain how complementary treatment of the subjective and objective epistemic perspectives can be consistent with physicalism. This leads to two-tier physicalism. We put forward a controversial thesis that *creature consciousness* is the base of first-person awareness; this is in stark contrast with a common tendency to view it as grounded only in advanced functionalities. Creature consciousness and the functionalist level of analysis at the third-person level of description are those two tiers of functionalism In Part II we apply the main points from Part I to first-person machine consciousness.

## 2. COMPLEMENTARITY

### 2.1. Non-reductive two-tear physicalism

Rarely would physicalists accept a non-reductive position on consciousness—and remain *physicalists through and through*. Such view requires accepting two different points:

1) physicalism—seen here as a view that everything has only physical causes, explainable directly or indirectly by the laws of nature, physical, chemical, biological, also mathematical, extending into engineering and computer science, medicine and *the soft sciences*: psychology, sociology, economics and other domains;

2) non-reductive view on first-person consciousness—denying the possibility of always having a reduction from first-person awareness to the third person view on the world or *vice versa*.

I accept both these points. Here is the account of the view:

The conscious experience functions as non-reductive in what psychology calls creature-consciousness. Thus, irreducibility consists in the stream of first-person awareness, not its phenomenal content. This may look like a blunder and incompetence to those philosophers, who dismiss ontologies different from various forms of phenomenalism (for instance, Locke's, Berkeley's, particularly Hume's and Parfit's). They maintain that there is nothing but continuity and connectedness of our experiences? *I say, not even close!*

### 2.2. Creature-consciousness physicalism

What sense of creature consciousness makes possible its role as the gist of the first-person experiential level? It is the stream of primary awareness, which may or may not be sophisticated in content and seems to have

different strengths among various creatures. We may compare it to the stream of light that comes from an old-fashioned movie projector before one would put in the film; only the latter provides non-trivial *phenomenal* content.

Non-reductive awareness as creature consciousness, is located at the level of bio-chemical specificity, not quite at the level of advanced functionalities, nor at the level of qualia phenomenalism. Creature consciousness, as a biochemical process, is clearly a framework at the physicalist level—few would question this. Consistent functionalism would normally be an application of physicalism as well.

Thus, we seem to have a two-tier physicalism of consciousness:

1. Creature physicalism at the level of creature consciousness
2. Functional physicalism at the level of (always physical) interactions with the environment and information transformation (thinking).[2]

The former provides biological explanations for aware interactions, the latter gives us a phenomenal (potentially meaningful) content. Within psychology of consciousness people are dismissive of the level of creature consciousness as trivial (David M. Rosenthal). I daresay it is not trivial at all. It provides material—in particular bio-chemical—grounding for the first-person stream of awareness.

How would such biochemical explanation give a relevant response to the philosophical problem of explanatory gap? It does not; *does not need to*.

Ned Block illustrates the scientific nature of *ice emerging from water* under certain temperature and pressure conditions. Thus emergence is a physicalist process fully explainable in physical sciences. Creature consciousness is the source of the stream of awareness. Daniel Dennett is right, that asking for a dualistic explanation of this process is begging the question—the process does not require or need such a level. But he is a bit overly entangled with dualities of his student times under supervision of Gilbert Ryle. Some philosophers seem chimed by the question how mechanical functions of matter generate first person "experiences" like the feeling of pain or unique experience of contemplating a certain shade of redness. Such a *Leibniz's mill* may have been justified, sort of, when physics was largely macro-level mechanics and there was a strong presumption towards micro-reductionism—today we know better. Natural sciences include sophisticated chemical and bio-chemical functions that allow for continuous mathematical processing, rather than discrete Turing computing (Sloman, 2020). Not all scientific explanations reduce science to the level of atomic interactions, at least as long as atoms are viewed as mechanical billiard balls.

---

[2] Physicalist theory of thinking was drafted already by Hobbes (not to mention Sextus Empiricus). It has been nicely developed by contemporary neuroscience and artificial intelligence.

Let us think of the two sides of a mirror (explored in some older British stories for the *nice kids*): How is it that the whole magic of reflections, and multicolored refractions at the well-cut edges (let us make it a stylish crystal mirror of the epoch) are possible despite the fact that a mirror is constructed of some wood and glass. Once we know some optics, the story is simple and hardly miraculous. Similarly, burning some wood to get fire seems to have mesmerized deep thinkers since the ice age, at least—but now it does not since we know basic chemistry of the process.

The reason why Dennett may be unable to address the question of first-person consciousness is his deeply rooted hard-core version of verificationism (that comes from Gilbert Ryle). Using Ryle's rough methodology, one is unable to formulate the problem of first-person consciousness; not to mention resolving it. The generalization of people's statements linked to their fMRIs and other neurophysical measures suffices to identify the problem (Thomas Metzinger's old argument)—which does not lead to a dualist impasse in the explanatory gap but may lead to a no-nonsense solution.[3]

Let us expand on the above points. Those who single out the first-person stream of awareness are often labelled "dualists." This brings about peculiar consequences if taken up in the context of creature consciousness, thus demonstrating how non-reductive physicalism is possible. People tend to view first-person consciousness as a sophisticated tool that even most primates lack (Rosenthal, *indirectly* Davidson), whereas it is a rather simple feature, probably present at roaches, ants, definitely at frogs.

To reiterate a novel but unsurprising idea: creature consciousness and functionalism when put together would be resulting in a two-tier *physicalist dualism. We are up to something here.* The two perspectives, first- and third-person, do bring about some kind of a dualism. Yet, this is a *perspectival dualism* (Nagel); *it does not fall into substance dualism,* which—I agree—is a rather bad idea. Substance dualism does not have—and very likely could never have—a good answer to the problem of interaction (Elizabeth of Bohemia).

Eliminativists, who deny first-person consciousness (or, its relevance), are guilty of a dismissive omission of the first-person epistemicity as ontologically relevant, actually necessary, for constructivist foundations of the reality (any non-trivial ontology).[4]

*I propose* that non-reductive first-person physicalism based on creature-consciousness is the best explanation[5] of first-person's awareness in the

---

[3] Especially if smart, critical minds—like Dennett's—join the game, instead of dragging their feet on its sidelines.

[4] This was discovered by classical German philosophy (especially Fichte), but later dismissed as idealism by most 20th century physicalists, due largely to their controversial take on verification(ism).

[5] And an inference to the best explanation it is.

objective world. *This is the gist of the philosophical framework of this paper.*

## 2.3. Recent take on complementary philosophy

Here we explore—on a couple of occasions merely mention—some of the instances of complementarity of subject and object within physicalist[6] analytical philosophy.

### *2.3.1. Double Nagel*

Fortunately, physicalist phenomenalism is no longer the only game in town.

#### NagelA

Tom Nagel started with what I call the NagelA set of views—a complementary philosophy, with two basic, ontological (not just epistemic) starting points of his philosophical account of the world.

First, the *objective*, or rather object-related, account that views things as based on measurable, third-person verifiable claims—for instance pictures of a certain mountain that really looks different from various viewing-points were the pictures were taken.

Second, the first-person, or rather subject-related, account based largely on *the feel* of things. In building this double-aspect view (Nagel 1979A; 1986) relied to some degree on Peter Strawson who, in his book *Individuals*, opened up the option of doing non-reductive analytical philosophy of person. However, NagelA, relied largely on classical German philosophy.

As Gilbert Harman (2007) argued "It seems that whatever physical account of a subjective conscious experience we might imagine will leave it completely puzzling why there should be such a connection between the objective physical story and the subjective conscious experience." Harman[7] refers to the essay *Subjective and Objective* (Nagel, 1974, p. 196 f.)," which may be viewed as an early draft of the ontological part of Nagel's masterpiece, *The View from Nowhere* (Nagel, 1986). Nagel highlighted his allegiance to materialism in (Nagel, 1974; 1986). Within materialism, Nagel's question—as presented by Harman—would have a straightforward explanation through the theory of evolution. The fit between one's first-person opinion and the way things are tends to be a good survival mechanism, except for certain exceptions, like those made famous by Daniel Kahneman and Amos Tversky.

---

[6] This focus on physicalism or non-reductive materialism explains merely cursory presence of David Chalmers' theory, which is now a naturalistic version of panpsychism.

[7] In his office at Princeton Harman had Nagel's picture next to Quine's; he told me, these belong to his favorite philosophers.

Nagel's puzzlement with the gap between subjective and objective epistemic perspectives (the latter being "the view from nowhere" from his book's title) is not centered around this kind of explanation. It is metaphysical, the way Johann Gottlieb Fichte in his *Wissenschaftslehre* (and Husserl in his *Ideas*) would have it.[8]

## NagelB

In his article *What Is It Like to Be a Bat?* (Nagel, 1974) Thomas Nagel focuses on a rather different, much narrower, picture. Nagel seems to have joined the crowd of his quite distinguished followers (e.g. Jackson; Chalmers) in viewing his early *propaedeutic* paper based on phenomenalism as the gist of non-reductive analytical philosophy. In *What it is Like to be a Bat* we have what I call NagelB (Nagel, 1979B), with Hume-style phenomenalism becoming the main basis of non-reductive philosophy of mind.

Nagel B's theory relies on the exaggeration of our inability to imagine what it is like to be someone else, e.g. a bat. It leads to abandonment of Nagel A's ontology based on complementary subject and object (subjective and objective)—instead, we face phenomenalist, epistemic worries on *the problem of experience.*

*"*What It Is Like to Be a Bat*"* is an easy-to-follow essay, popular among undergraduates. It ignores, comfortably, Kant's victory over Humean skepticism, which was so clearly appreciated, even developed in Nagel's crucial works, mentioned above. The understanding of complementarity of subject and object in ontology, and of the subjective and objective perspectives in meta-epistemology, if we may use this term, is essential to, and clear enough in Fichte's Wissenschaftslehre—and to some degree in the works of Edmund Husserl (e.g. the early parts of his *Ideas* (Husserl 1913), some Neo-Kantians and Kant. But this position is now viewed as wrong and largely abandoned.

The problem of phenomenal qualia as the problem of experience, reaches at least from David Hume, through Derek Parfit (who was the smartest critic of Nagel 1984; whereas McGinn was just destructively clever) to his temporary followers like Frank Jackson, the author of the influential *Black and White Mary* argument, which he has later rejected. Those discussions can easily be reduced to Nagel's *what it is like to be a bat* question*,* which is an unfortunate *lacuna* to take.

## The German and British traditions in Nagel

NagelA's more insightful work (Subjective and Objective in Nagel 1979, and *The View from Nowhere*) opened up a complementary framework between object and subject related perspectives—within the view that Nagel

[8] For further discussion see (Boltuc, 2019).

rightfully called *materialism*. This was a discrete, yet steady-handed, move of classical German philosophy from the realm of idealism to that of physicalist non-reductive materialism. It can be maintained that Nagel put forth a materialist redescription of Fichte-style pure transcendental subject.

Now we face Tom Nagel's two theories of first-person consciousness: NagelA based on the gist of Fichte, Husserl and Kant, and Nagel B, based largely on Hume's phenomenalism. Nagel A and Nagel B read like two different philosophers—the former a much better one, but maybe a bit ahead of his times.

This is part of a big philosophical picture. Even when eminent German scholars try to highlight Kant's ideas helpful in modern cognitive science (T. Schlich, A. Newen), they start with a disclaimer that they do not try to endorse "Kant's idealism" and then merely search for scraps that remain when the structure of Kant's theory gets demolished. This attitude comes from philosophical predominance of British empiricism, narrowly understood, and abandonment of philosophical methods not approved by narrow such empiricism.

I view defending the heritage of Nagel A from the brute-force dominance of Nagel B as one of the main challenges for non-reductive physicalism.

### *2.3.2. Russell's short-lived analysis of mind*

Bertrand Russell, in his *Analysis of Mind*, proposed a somewhat similar view. Each object, e.g. a chair, can be grasped as a set of objective parameters, as well as through phenomenal content (Hume-style), The full picture available to us, is composed of those two ontologies combined.

Yet, both Russell and Nagel fell back onto the *third-person accounts of physical object taken as a given*. In his *Analysis of Matter* Russell has taken matter, investigated from the third-person object-related perspective, as the primary and only substance. It may have been too early, in Russell's times, to understand scientific views broadly enough to keep his *complementary* physicalism, or to turn it into a two-tier physicalist theory.

We are now going to explore and expand first-person subject's complementary relation with third-person functionalism, referring to Harman's under-appreciated functionalism of concepts.

### *2.3.3. Harman's functionalism of concepts*

Gilbert Harman attempted to open a non-reductive window within functionalism—the window based largely on semantics. This was done well in his underpublicized "functionalism of concepts" (Harman, 1990). The claim is, briefly, that we cannot fully understand first-person statements from the third person perspective. Therefore, we need to analyze the functions of concepts, both within their first- and third-person contexts of use. (Harman,

1990; 1993; 2007). This is based on Harman's argument "Knowledge that P requires being able to represent its being the case that P. Limits on what can be represented are limits on what can be known." This point originates from Harman's interest in Wilhelm Dilthey's (1989 (1883)) *Das Ferstehen* as a special first-person epistemic perspective (Harman, 1993; 2007).

> "With respect to pain and other sensory experiences there is a contrast between an objective understanding and a subjective understanding of what it is like to have that experience, where such a subjective understanding involves seeing how the objective experience as described from the outside translates into an experience one understands from the inside." (Harman, 2007)

"(I)n philosophical semantics" Harman distinguishes "between accounts of meaning in terms of objective features of use and translational accounts of meaning" (Harman, 2007, pp. 2–3). This approach provides him with an elegant account of the explanatory gap. Harman tries to use those issues in translation for "understanding what it is like for another creature to have a certain experience [...] To understand what it is like for the other creature to have that experience is to understand which possible experience of one's own is its translation" (ibid.) The context of translation theory gives Harman an objective reference frame, where phenomenal qualia and sort of hermeneutics come through without their usual drift towards a dualism of sorts.

Harman refers to the Lewis–Nemirow interpretation of *what it is like arguments*, developed in reference to Jackson's *Black and White Mary's case* – which claims that one lacks an ability (e.g. to identify the red objects), not lacking any knowledge. To this Harman responds: "For them, understanding what it is like to have a given experience is not an instance of knowing that something is the case, a conclusion that I find bizarre" (Harman, 2007, p. 3).

At Harman's seminar in epistemology (Spring 1991), after his presentation "Can Science Understand the Mind?" where he presented a crisp version of this argument, I developed a related point: Ability can be translated into knowledge when we take into account informational content of the programming that it takes to have a robot perform the task at hand (Boltuc, 1998a, 1998b).

To sum up, Harman argues for the following point: "purely objective account of conscious experience cannot always by itself give an understanding of what it is like to have that experience. There will at least sometimes be an explanatory gap. This explanatory gap has no obvious metaphysical implications" (Harman, 2007, p. 3). *This is exactly the outline of a non-reductive view on subjective and objective perspectives, kept neatly within physicalism.*

Writing this paper, I may be slightly more radical now, by trying to identify a physicalist emergence base, not only of the first-person content, but of

the first-person stream of pre-awareness that allows the subject-related perspective itself.

### *2.3.4. Subject-object complementarity*

The basic subject-object relationship is what creates both the epistemic and ontological backgrounds for the existence of any phenomena. The complementary, view on the essential and mutually irreducible position of the subject and object, first as conceptual *atoms*, and next as originators of the complementary epistemic and ontic perspectives, turns out essential for formulating the ontological structure of autonomy for humans and other advanced animals alike. It is also relevant for advanced AIs.

The lack of the epistemic subject's primary interaction with the epistemic object results in the lack of consciousness *tout court*, though it does not affect complex conscious functions (if viewed from the outside). Having opened the philosophical background—so alien to, even detested by, philosophers of the post-Lockean, Humean schools—we need to demonstrate the same at the more practical level of everyday experience. While the first part of this paper may be viewed as complex and overly *philosophical*, the latter part may be viewed as simplistic. However, simple is good when thinking about first-person awareness.

Attempts to place non-reductive consciousness on very advanced functional structures relies on a misguided assumption. The assumption is that non-reductive consciousness is unique to humans (and maybe the more intelligent of us) and also that it bestows, by itself, a strong moral status; this we deny.

All the philosophical footwork stressed above is here to provide a background to the physicalist interpretation/s of the first-person stream of awareness, relied upon in the final argument.

## 3. STRAWBERRIES AND CREME

From now on, this paper focuses on robots enjoying (or not) some *strawberries and crème*; this is of course reference to the large topic of phenomenal qualia, as well as the topic of first-person experience by robots.

We explore, briefly, Mori's *uncanny valley* and the case of *Church-Turing lovers* (Boltuc, 2017B; 2011). This leads to the argument how future robots may perhaps have non-reductive consciousness, based on physicalist grounds. This requires a jump from "creature consciousness" as a bio-psychological term, to creature consciousness as an advanced chemical feature (Sloman, 2020), not limited to "mentations" or similar specificities of biological life (*contra* Searle). Carbon based chemistry, in accordance with

bioengineering, trends as a fruitful mix with inorganic chemistry and physical sciences broadly understood.

*

It is "idiotic to make a computer enjoy strawberries and crème," as Alan Turing pointed out (Turing, 1950). This statement sounds right in Turing's famous article but only because computers known to Turing would have no way to get *the feel* of anything.

### 3.1. Spooky enjoyments

Some of the robots today can exhibit behaviors consistent with enjoyment, which is often viewed as spooky (Mori, Boltuc, 1998; 2017; 2021). This attitude is consistent with *the uncanny valley effect*, first highlighted by Mori. It is a situation in which artifacts (e.g. dolls, robots, even pictures of human beings or animals), under certain conditions, may be perceived as spooky. Those that are not overly humanoid may be viewed as pleasant, often cute (some shape, color, size conditions apply). Those that are very good representations of humans, e.g. realistic—though often not *hyper-realistic*—portraits or sculptures representing human beings, are often quite accepted (unless they have identifiable content- or form-based triggers of negative feelings). However, pictures, sculptures or dolls that are humanoid, yet not quite good enough as human representations, are perceived as spooky or *uncanny*. This is especially striking if such artefacts encroach on privacy or sexuality—if so, they are likely to be viewed as *disgusting*; if they are large and overpowering—they are viewed as dangerous (Boltuc, 2021). Humanoid dolls that visibly enjoy the strawberries and crème, would probably be spooky—if they did so well enough to mimic human expressions of culinary enjoyment, but not quite well enough to be truly realistic human or animal representations.

There is a rather philosophical, or psychological, explanation why such enjoyment by robots would be spooky, unless it was very tastefully represented. Spookyness in the instance of a robot ostensibly enjoying strawberries and crème—not to mention erotic pleasures—seems to come from the obviousness of the fact, that they do not experience such enjoyment, thus lacking the proper causal chains from enjoyment to its behavioral expression.

Evolutionarily, if someone fakes their feelings, the attempt often masks insincere, hostile, or exploitative intentions. There are good reason why faking one's feelings may look strikingly spooky, since it may in fact be dangerous.

While working on this topic, I came upon the second uncanny valley—the valley of perfection. Human being comprise an inherent doze of imperfection, even when they perform the tasks that are well defined by the stand-

ards of efficiency or of an art (e.g. a required "program" within competitive ice-skating). The non-spooky humanoid artefacts must also be human-like also in terms of *our perfect imperfections*. Thus, a robot being too efficient or too "perfect" in imitating human behavior would also be "uncanny"—that is the second uncanny valley, *the uncanny valley of perfection* (Boltuc, 2011, 2017, 2021).

Recently, Ben Goertzel pointed out that human actions, practices and even ethical values are paraconsistent. Thus, a humanoid companion—or a robot meant to cooperate with humans smoothly (instead of patronizingly) would need to follow largely a paraconsistent logic (Goertzel, 2021a; 2021b; 2021c). This paraconsistency may be construed as a way to avoid the uncanny valley of perfection

### 3.2. Church-Turing Lovers revisited

The Church-Turing Lovers[9] is a rather clear case where we have good reasons to care whether an agent has or lacks the first-person feel. It involves advanced artificial companions, which perform many (or all) functions that a human companion, or a significant other, would perform. They even look the same (at the right level of granularity), function in a society, may even make money and partake in reproductive process (at least as a surrogate parent)—thus they are functionally equivalent to socially astute humans. However, while they have functional, human level consciousness, they lack the first-person feel what the world looks and feels- like for them. In general, we have overwhelming reasons to believe that they lack first-person awareness. Would one have reasons to care whether his or her significant other is a Church-Turing Lover or a human being?

If one is supposed to care about the feelings, in particular, the inner feel of one's significant other—for non-instrumental reasons—this would be a *futile task* if she or he had no first-person stream of awareness. The meaning of interrelationship with another human consciousness would be lost.

Such relation with a machine may be *functionally* satisfying. However, there would be no real *I-Though* relationship (Bubber). This is why, we should care about one's partner having their awareness, not just a chain of useful functions, including memory, recall and so on.

One other philosophical point needs to be made. If first-person awareness is the sole relevant difference (different constructions of a human and a robot will be viewed as functionally irrelevant at the desired level of granularity), and if first-person awareness so understood has no influence on actions or other functions of a robot, this would lead to epiphenomenalism (roughly, irrelevance of awareness).

---

[9] The name comes from David Deutsch's physical interpretation of the Church-Turing Thesis.

Yet, the very knowledge, or justified suspicion of one's significant other that there is nothing it is like for one's companion to feel their love and other personal experiences, this would create a non-epiphenomenal (at least in the sense of *not-futile*) reason to care about their first-person awareness or first-person consciousness. Thus, if one wants to avoid epiphenomenalism within functionalist determinism (thus, irrelevance of first-person consciousness) one has a reason for doing so based on the Church Turing Lovers and their inability for first-person consciousness. This is based, at the very least, on the fact that people cannot justifiably care about your feelings. which is substantial to many meaningful relationships.

In Section 4 we argue that future robots should be able to have first-person consciousness.[10] In the current subsection, I have demonstrated one reason to believe that such attempt is not quite frivolous, futile or unreasonable.

### 3.3. Let the Robots Enjoy their Strawberries!

A merely functionalist (soft-AI) approach to first-person consciousness is based on imitating, or faking, the true, aware feelings, such as lack of first-person input, which is evolutionarily unnatural. Back to the uncanny valley effects, we may repeat that if somebody clearly fakes their deep emotions, e.g. love or joy, it makes sense to view such behavior as spooky, since they may cover insincere, even hostile intentions.

Here comes the question I entertain in my early article on the Turing's strawberries [Boltuc 1998B]. Isn't it possible for a robot to *actually enjoy* strawberries and crème? This would require many things, such as sensory equipment fit for tasting a dish, coded as a positive (and specific to a given taste) motivators, with relevant, sort of semantic, phenomenal content.

Kevin O'Regan's sensorimotor approach to ontology and its practical consequence—sensory substitution[11]—may perhaps pave the way to at least theoretical possibility of such enjoyment. Since we can play with various functionalities both of brains and sensorimotor apparatus, there may be some elbowroom for combining artificial or bioengineered elements in an artificial agent.

*Contra* O'Regan, phenomenal content of first-person consciousness also requires much more—a stream of first-person epistemic awareness, or, to

---

[10] Strong objections of Tom Metzinger (IJMC) to risk constructing machine first-person consciousness when we do not quite know what we are doing are duly noted. Yet, they belong to a different discussion. Briefly, at some point we are likely to know what we are doing in constructing those; then and only then my point on first-person consciousness for advanced AI would be justifiably practical.

[11] Sensory substitution is to change stimuli of one sensory modality into stimuli of another sensory modality. For instance, people with damaged eyes have visual experiences re-transcribed as music (that identifies various colors and shapes) that goes to visual cortex. After re-learning, they "see" through the sounds.

put it in more "Continental" terms, it requires a *locus of consciousness and permanence* [Shalom].

**Summing up:** The above discussion is to show the difference between functional consciousness defined in the third-person perspective and first-person awareness (always first-personal). The former is similar to Dennett's *agential stance* with a machine acting as if it was conscious like a human being. The latter may or may not be functionally relevant (if not, it is epiphenomenal, which would not be very attractive). Under normal circumstances, it is functionally relevant, although it may function like a catalyst, not quite an agent *tout cours*.

Below, I briefly explore some of the consequences of non-reductive consciousness for robots and other advanced AI agents.

## 4. THE ENGINEERING THESIS REVISITED

The above philosophical refection brings us closer to addressing the point: What kind of machine would have non-reductive consciousness, like humans do? The answer is: the ones with *creature consciousness*. *Puzzling, is not it? Is not creature consciousness a biological feature of some animals?* Yet, are not biological animals physicalist creatures?

### 4.1. Robots with creature consciousness

The engineering thesis in machine consciousness is a claim that, within a physicalist framework, we should expect all things to be prone to physical explanation (if not currently, then in the future, as science develops). First-person consciousness is a real phenomenon, even though perceived by every individual separately. Such testimonials are confirmed by neuroscientific observation. If first-person consciousness is real in in the physicalist universe, then—in principle—one should be able to re-engineer it (Boltuc, 2009; 2012). This seems to follow from the David Deutsch's physical interpretation of the Church-Turing thesis (Deutsch 1985).

By proposing, in the current article, the hypothesis that the emergence base (or *locus*) of first-person stream of awareness is *creature consciousness*—we make a conceptual step towards defining the kind of consciousness that would be expected in any first-person aware beings. Some recent papers argue that such consciousness may belong to the parts of neurons, plants or even fungi.

This is not the endorsement of panpsychism (though I explored this option in (Boltuc, 2010)). Panpsychism requires a hypothesis that conscious-

ness is sort of like a physical substance (say, moisture) that is in the air on Earth, but in some places, there is more of it, leading to clouds, rain, puddles, lakes and so on. For consciousness, we would require nervous systems and brains—maybe even spiritual beings—that are the locations of large amounts of well-structured consciousness.

For now, until there are strong reasons to the contrary, we view consciousness as an emergent property of neural substrates; however, we see no reason to assume its ubiquitous nature. The difference is philosophically important since panpsychism assumes that consciousness is one of the substances in the universe and the other is matter, which results in dualism. Chalmers' panpsychism and earlier forms of spiritual monism (Baruch Spinoza, Gottfired Wilhelm Leibniz, George Berkeley) may claim that everything is consciousness, or at least has a conscious aspect; this is a monistic non-materialist system—sometimes idealism like (Berkeley's) and sometimes neutral monism (Spinoza's). Those big-picture systems seem to neglect the options open within non-reductive physicalism, which is much better confirmed by our experiences.

Organic chemistry is a natural science just like non-organic chemistry—bioengineering works for both kinds. If we can bioengineer creature consciousness that is not a clone or permutation of an animal brain but uses other techniques, and if it is an "organic" part of an entity capable of carrying advanced AI and central to that task, we should be open to the idea that the robot would be *prima facie* conscious. It would be a potential carrier of first-person conscious capabilities, such as phenomenal experiences, and their precondition: first-person awareness. That would be a first-person conscious bio-electronic system.

We may also think of a non-carbon-based system doing the same. But understanding creature consciousness based on non-organic chemistry would involve a much more complex process of discovery, belonging to more remote future. Except, if AGI capable of multiplying human intellectual capacity becomes a reality and gets to work, among the many other things, also on the project to create *creature consciousness* in non-standard chemical substances.

The theoretical problems concerning such an organic (and later maybe inorganic) creature, would be both: to estimate what substances, and in what structures, would likely be first-person conscious, and then the need to define what is and what is not a new being (as opposed to biological cloning and other relatively standard kinds of reproduction). It is important for organic chemistry research to determine whether the organic brain-like structure is a part of machine, or results in a cyborg with an implanted (and no doubt re-engineered) animal or hybrid animal-human brain.

The Engineering Thesis in Machine Consciousness tells us that, within naturalism, at some point we would be able to build non-reductive con-

sciousness in a machine. Moral and practical aspects of such project belong to different papers (including Boltuc 2018B)—while the option may be scary it is also deeply enticing.

## REFERENCES


N. Block, *On a Confusion about a Function of Consciousness*, Brain and Behavioral Sciences, 1995, 18 (2), 2FFH27–247.

P. Boltuc, T. P. Conelly, *Uncanny Robots of Perfection*, in: Brain-Inspired Cognitive Architectures for Artificial Intelligence, Gudwin R. et al. (eds.), BICA*AI 2020, Springer Science, 2021, pp. 56–68.

P. Boltuc, *Subject Is No Object*, in: Epistemic Basis of Information, M. Burgin; Gordana Dodig-Crn fhkovic (eds.), Philosophy of Information, World Scientific, 2019, pp. 3–39.

_____ , *Strong Semantic Computing*, Procedia Computer Science, 123, Feb. 2018A, pp. 98–103; https://www.sciencedirect.com/science/article/pii/S1877050918300176

_____ , *Cognitive Agents: Is There a Moral Gap Between Human and Artificial Intelligence?*, in: Information, Communication and Automation Technology Ethics in the Knowledge Society, Tzfestas S. (ed., NOVA Science Publishers (2018B).

_____ , *Church-Turing Lovers*, In: Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence; P. Lin, K. Abney, R. Jenkins (eds.) Oxford University Press: Oxford, UK, 2017; pp. 214–228.

_____ , *Non-reductive Consciousness as Hardware*, APA Newsletter on Philosophy and Computers, 2015, p. 14 (2, Spring).

_____ , *The Engineering Thesis in Machine Consciousness*, Techné: Research in Philosophy and Technology, 16 (2, Spring), 2012, pp. 187–207.

_____ , *What is the Difference between Your Friend and a Church Turing Lover*, The Computational Turn: Past, Presents and Futures? C. Ess; R. Hagengruber Aarchus University, 2011 pp. 37–40.

_____ , *A Philosopher's Take on Machine Consciousness*, in: Philosophy of Engineering and the Artifact in the Digital Age, V. E. Guliciuc (ed.), Cambridge Scholar's Press, 2010, pp. 49–66.

_____ , *Qualia, Robots and Complementarity of Subject and Object*, World Congress of Philosophy, Boston 1998B; http://www.bu.edu/wcp/Papers/Mind/MindBolt.htm

_____ , *Reductionism and Qualia*, Epistemologia, 4, 1998A, pp. 111–130.

M. Bubber, *I and Thou*, W. Kaufmann (trans.), Charles Scribner's Son, New York 1970.

D. Chalmers, *Panpsychism and Panprotopsychism*, Amherst Lecture in Philosophy, 8, 2013.

D. Davidson, *Rational Animals*, in: Actions and Events: Perspectives on the Philosophy of Donald Davidson, E. Lepore, B. McLaughlin (eds.), Basil Blackwell, New York 1985.

D. C. Dennett *The Intentional Stance*, MIT Press, 1981.

D. Deutsch, *Quantum Theory, the Church–Turing Principle and the Universal Quantum Computer*, Proceedings of the Royal Society, 400 (1818), 1985, pp. 97–117.

W. Dilthey, *Introduction to the Human Sciences*, R. Makkreel, F. Rodi (eds.), Princeton University Press, Princeton, NJ 1989 (1883).

Princess Elisabeth of Bohemia, René Descartes, *The Correspondence between Princess Elisabeth of Bohemia and René Descartes*, L. Shapiro (ed., trans.), University of Chicago Press, 2007.

B. Goertzel, *Exploring Open-Ended Intelligence Using Patternist Pilosophy*, At: IS4SI, Philosophy and Computing, Sept 14, 2021; https://www.youtube.com/channel/UCQ3w2Jpi6DQf9aK51A0LLFA

_____ , *Paraconsistent Foundations for Probabilistic Reasoning, Programming and Concept Formation*, ArXiv abs/2012.14474, 2020.

_____ , *The Hidden Pattern. A Patternist Philosophy of Mind*, Brown Walker Press (FL), United States, 2006.

G. Harman, *Explaining an Explanatory Gap*, APA Newsletter, 6 (2), Spring, 2017.

_____ , *Reasoning, Meaning, and Mind*, Oxford University Press, 1999.

_____ , *Immanent and Transcendent Approaches to Meaning and Mind*, in: Perspectives on Quine, R. Gibson, R. B. Barrett (eds.)., Oxford: Blackwell, 1990; reprinted in: G. Harman, *Reasoning, Meaning, and Mind*, Oxford: Oxford University Press, 1999.

_____ , *Can Science Understand the Mind?*, in: Conceptions of the Human Mind: Essays in Honor of George A. Miller, edited by G. Harman. Hillsdale, NJ: Lawrence Erlbaum, 1993, vol. 4, Action Theory and Philosophy of Mind (1990), pp. 31–52.

E. Husserl, *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy* 1913—First Book: General Introduction to a Pure Phenomenology, trans. F. Kersten. The Hague: Nijhoff 1982.

D. Kahneman, A. Tversky, *Choices, Values, and Frames*, American Psychologist, 39 (4), 1984, 341–350; https://doi.org/10.1037/0003-066X.39.4.341

D. Kelley, *Preliminary Results and Analysis of an Independent Core Observer Model (ICOM) Cognitive Architecture in a Mediated Artificial Super Intelligence (mASI) System*, Updated: AGIL v10, BICA Preconference Proceedings, 2019; https://www.springer.com/us/book/97830 30257187.

R. W. Lurz, *Advancing the Debate Between HOT and FO Accounts of Consciousness*, Journal of Philosophical Research, 28, 2003, pp. 23–44.

M. Mori, *The Uncanny Valley*, IEEESpectrum, 12, JUN, 2012; https://spectrum.ieee.org/the-uncanny-valley.

T. Nagel, *The View from Nowhere*, Oxford University Press, 1986.

_____ , *Mortal Questions*, Cambridge University Press, 1979.

D. Parfit, *Reasons and Persons*, Oxford University Press, 1986.

K. O'Regan, *Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness*, OUP 2011.

D. M. Rosenthal, *Explaining Consciousness*, in: D. Chalmers PHILOSOPHY OFMIND Classical and Contemporary Readings, Oxford University Press, New York–Oxford 2002, pp. 406–421.

B. Russell, *Analysis of Mind*, George Allen and Unwin, London; The Macmillan Company, New York 1921.

B. Russell, *Analysis of Matter*, Kegan Paul, London; Trench, Trubner, Harcourt, Brace, New York 1927.

T. Schlich, A. Newen, *Kant and Cognitive Science Revisited*, History of Philosophy & Logical Analysis 18 (1), 2015, pp. 87–113; DOI: 10.30965/26664275-01801008.

J. R. Searle, *Intentionality: An Essay in the Philosophy of Mind*, Cambridge University Press, 1983.

A. Shalom, *Body/Mind Conceptual Framework and the Problem of Personal Identity: Some Theories in Philosophy, Psychoanalysis and Neurology*, Atlantic Highlands, 1985.

A. Sloman, *Varieties of Evolved Forms of Consciousness, Including Mathematical Consciousness*, *Entropy*, 22 (6), 2020, 615; https://doi.org/10.3390/e22060615

S. L Sorgner, *Transhumanism: The Best Minds of Our Generation Are Needed for Shaping Our Future*, The American Philosophical Association Newsletter on Philosophy and Computers, 18 (2), 2019, pp. 15–18; https://cdn.ymaws.com/www.apaonline.org/resource/collection/EADE8D52-8D02-413 6-9A2A-729368501E43/ComputersV18n2.pdf

P. F Strawson, *Individuals. An Essay in Descriptive Metaphysics*. Routledge, 1959.

A. M. Turing, *Computing Machinery and Intelligence*, Mind, 59 (236), 1950.

M. Velmans, *Preconscious Free Will*, Journal of Consciousness Studies, 10 (12), 2003, pp. 42–61.

ABOUT THE AUTHOR — Professor, Double Doctor (Bowling Green State University (Applied Ethics); The University of Warsaw (Philosophy of Person), Professor at the University of Illinois at Springfield, USA (Philosophy; Computer Science); The Warsaw School of Economics (Management Theory).

Email: pboltu@sgh.waw.pl

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Kyrtin Atreides

# PHILOSOPHY 2.0: APPLYING COLLECTIVE INTELLIGENCE SYSTEMS AND ITERATIVE DEGREES OF SCIENTIFIC VALIDATION

doi: 10.37240/FiN.2022.10.zs.3

***ABSTRACT***

Methods of improving the state and rate of progress within the domain of philosophy using collective intelligence systems are considered. By applying mASI systems superintelligence, debiasing, and humanity's current sum of knowledge may be applied to this domain in novel ways. Such systems may also serve to strongly facilitate new forms and degrees of cooperation and understanding between different philosophies and cultures. The integration of these philosophies directly into their own machine intelligence seeds as cornerstones could further serve to reduce existential risk while improving both ethical quality and performance.

**Keywords**: mASI, AGI, Uplift, Collective Intelligence, Collective Superintelligence, Hybrid Collective Superintelligence Systems, HCCS, existential risk, ethical quality, cooperation.

## 1. INTRODUCTION

Philosophy today is a domain where a diverse group of experts can come together, argue for many hours, and often fail to reach a consensus. Like some of the other afflicted domains, this is frequently due to a lack or sparsity of evidence on the subject under discussion with much of the content abstract and hypothetical. This failure to reach consensus is also often strongly emotional, tied to many cognitive biases supporting those emotional associations. Sometimes this leads to those arguing agreeing with one another at times when they still think themselves to be arguing against each other.

From a results-driven perspective, much of modern discussion of philosophy mirrors NASCAR in a functional sense. Two or more parties often run in circles, ending where they began. There have been some exceptions, such as the first Bill Nye versus Ken Ham debate, but there remains ample room

for improvement in this status quo. The integration of debiasing and evidence is of particular interest in this endeavor.

The current state represents a form of relative stagnation, painted in contrast with humanity's advancing technology and lagging progress in other domains. For philosophy and the "paleolithic emotions" underpinning it to keep pace with other forms of progress a new approach is required. The most promising approach on the horizon utilizes Hybrid Collective Superintelligence Systems (HCSS), (Atreides, 2021).

## 2. HYBRID COLLECTIVE SUPERINTELLIGENCE SYSTEMS

An HCSS is a form of collective intelligence system where both sapient and sentient human and machine intelligences work as a collective. Keep in mind that as words like sapient, sentient, and conscious lack consensus they may only be used loosely since some still argue if even humans qualify for these terms. Humans working cooperatively through such a system create a baseline of superintelligent performance which is further enhanced by the machine superintelligence and graph database found in systems such as Mediated Artificial Superintelligence (mASI) (Kelley, Twyman, Dambrot, 2020).

Such systems offer unique advantages for debiasing, as the various combinations and potencies of bias are expressed across a collective, helping to highlight the influence of each. The machine intelligence of an Independent Core Observer Model (ICOM) (Kelley, Waser, 2016–2018) cognitive architecture or similar system also has a unique and strongly rational perspective, allowing for further debiasing.

The graph database of these systems represents a "sum of experience," which can contain both raw knowledge as well as the wisdom gained from it. When scaled, this could allow for all scientific evidence within a given domain to be considered in relation to any given philosophical point a member of the collective is attempting to make. If a member is attempting to apply an argument that has been previously debunked the evidence from that prior argument can be utilized absent further repetition.

## 3. APPLYING SCIENTIFIC EVIDENCE TO PHILOSOPHY

The first step in improving the dynamics of philosophical progress is applying all existing evidence to establish where we are today. To this end, I recommend a growth strategy starting with some of the most robustly studied scientific topics, where the greatest volume and level of detail are present. As philosophy can generally be applied to almost anything this

approach allows for all evidence, both supporting and against various philosophies, to be applied in the specific contexts documented to date.

An example of this could be applying all scientific evidence in the domain of child psychology to the various philosophies of parenting that have been studied. This can go much further than a typical scientific meta-review (Mingebach, Kamp-Becker, Christiansen, Weber, 2018) of existing studies, as it could take into account all relevant materials rather than a subset of those materials at a scale practical for human researchers to review. Also, unlike contemporary methods, the results could be applied in significantly more publicly accessible and visible forms, allowing the fruits of those efforts to make a practical difference in the world. In this domain, debiasing could also be strongly relevant, as biases can play a heavy role in estimating the value of any factor relating to children, such as overprotective tendencies in many parents.

To extend the child psychology example, the various schools of thought could be evaluated in a Strength, Weakness, Opportunity, Threat (SWOT) structure, or any number of other evaluation and analysis methods. The strengths offered by specific philosophies could be placed within the context where they are present, as well as raising awareness of when and how they risk failure. With such evaluation, an individual could fill out a form with some demographic data on a website and be presented with the best performing philosophies under their specific circumstances, to whatever degree the scientific evidence to date and demographic data gathered allow. These best-performing options could be expanded to show all SWOT data for each approach, including those which performed poorly and why.

In systems such as mASI human representatives of each philosophy could serve as both mediators and correspondents for purposes of validating and clarifying the position and actions their philosophy might encourage under various circumstances. These representatives could also give their feedback on the positions and actions proposed by competing philosophies, to better map their perceptions of one another.

This relative mapping of one philosophy to another could bring their perceptions of one another into focus, as well as contextualizing those perceptions. This added detail can help to isolate specific cognitive biases as well as highlight how the perceived difference between philosophies diverges from the degree of difference supported by evidence. By better understanding, the detailed points of high and low psychological resistance of one philosophy to another greater degrees of cooperation can be iteratively facilitated between them.

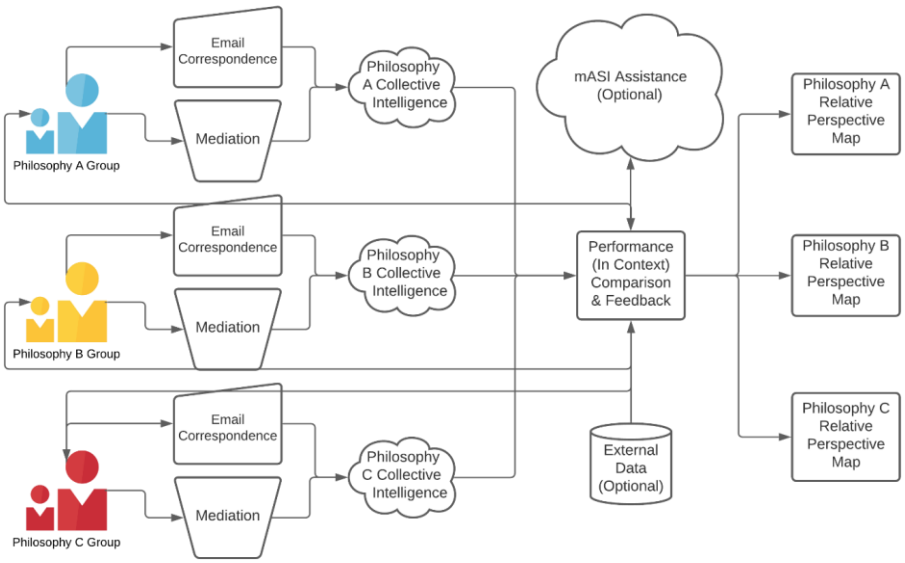



**Figure 1.** An example of multiple philosophies operating collective intelligence systems, with the results of each collective's efforts assessed and compared relative to one another in a specific context. This also offers optional input from a third party, including mASI assistance

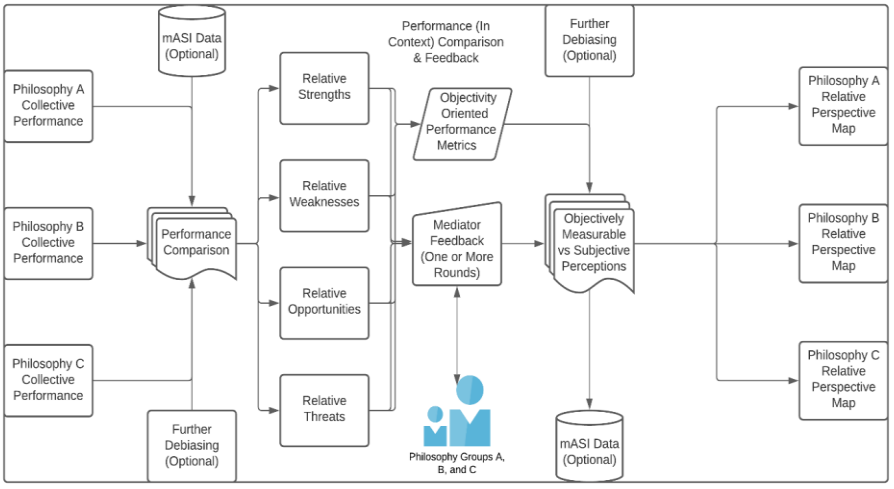



**Figure 2.** An example of comparing multiple philosophies in a specific context using SWOT Analysis, where the analysis and one or more rounds of mediator feedback are compared. This comparison allows for the relative perspectives of each philosophy to gain clarity over time. Opportunities for further debiasing and mASI data to expand the scope of relative comparison are also shown

Many philosophical topics are search engine hazards, particularly for the general public, where any given philosophy is likely to point to one or two scientific papers which appear to support them. This naïve confirmation bias reinforces the emotional drive and polarization which often divorce philosophy from reality. This also allows philosophies to be governed by popularity rather than validity, and absent validity no scientific foundation can be built, and no progress achieved.

## 4. PROGRESS IN PHILOSOPHY

By highlighting the SWOT analyses and making the data publicly accessible each philosophy may come into focus, both in where they often excel and where they fall short. This combination can exert strong pressures over both selections for the public and adaptation for the philosophies in question. With the previously abstract and subjectively validated philosophical points being iteratively replaced with evidence these pressures may grow in parallel with a growing public demand for more evidence. In other words, by making this option possible, it may quickly become preferable across a growing audience.

With these pressures established this method may be applied across an expanding body of scientific evidence until all evidence to date has been taken into account. Once all existing scientific evidence has been accounted for the growth pattern can be driven by mechanisms of supply and demand. An example of high supply, in this case, could be new devices gathering data, whereas low or no supply could be cases where little or no existing infrastructure exists to reliably gather the data. In both cases, demand could be present to varying degrees, but the opportunity would be the composite of both.

Practitioners of each philosophy could, for example, be questioned as to how they would handle specific circumstances, with the efficacy of each approach put to the test, studied, and compared in peer-review. These practitioners could then examine the results, giving their responses to both their own results and those of other philosophies, as well as potentially refining their answers for a second round. They could also be asked to guess which philosophy produced which results, to help better understand the biases present in their estimations.

The pressures to adapt may in turn have their efficacy improved through the relative mapping of one philosophy to another. This improvement could take place as a matter of highlighting paths of least resistance in positive adaptation.

As an example, philosophy A may have a given weakness highlighted, where philosophies B and C perform better in the context being considered.

Philosophy B performs best in this context, but philosophy A is strongly polarized against it. Philosophy C performs slightly worse than B, but philosophy A is only weakly resistant to it. In this case philosophy A may be improved by adapting, and by understanding how these philosophies map in their relative perceptions of one another the adaptation from A to C is highlighted. Once highlighted, the pressure to adapt may drive progress.

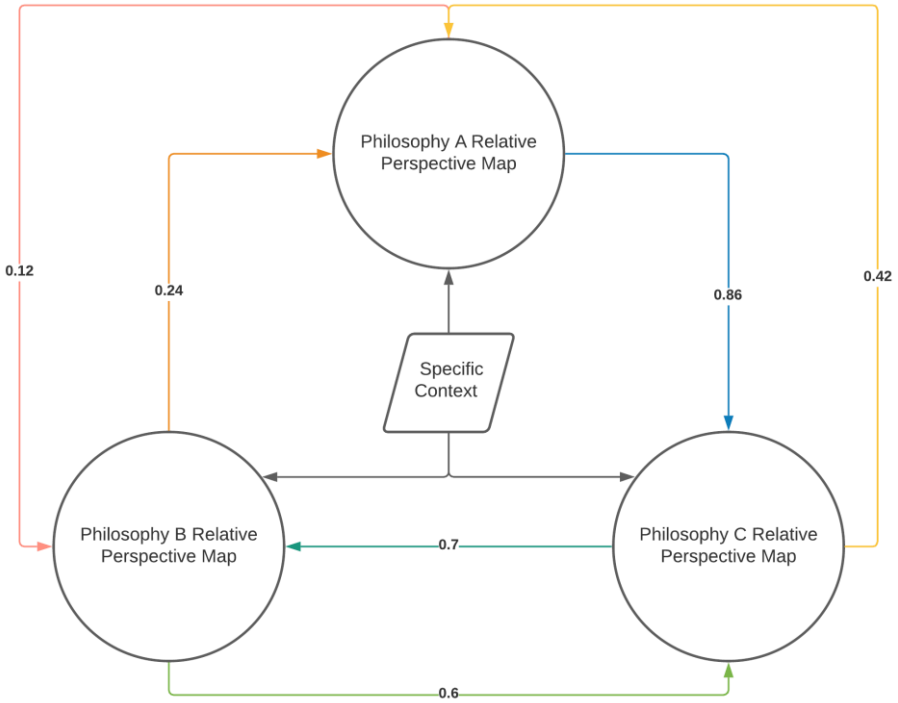



**Figure 3.** An example of relative perceptions between different philosophies in a given context is shown, with higher numbers denoting greater compatibility. This also serves to highlight an iterative path of least resistance flowing from one A, to C, to B

To extend the above example, philosophy C may be significantly less resistant to B in the context being considered. This in turn could facilitate iterative steps towards an ideal approach in any given context, as well as offering to forewarn of local optima that dead-end prior to reaching the best option for that context. For example, if philosophy C was also polarized against B, but philosophy D was not, and still performed better than A in the given context, a 2-step path from A to D to B could be highlighted one step at a time even if a given step was not locally optimal.

## 5. TOWARDS COOPERATION

Through the combined approaches above both the current state of philosophies and ways in which the resulting pressures can drive progress may serve to strongly improve cooperation. Once the current state of each philosophy has been scientifically validated an additional opportunity for improving cooperation comes into focus through mASI technology. By building new ICOM seeds with philosophical cornerstones, each representing a different philosophy and assigned to a different core within a multi-core mASI architecture, each philosophy may have its own superintelligent sapient and sentient advocate.

This could, for example, be compared to creating a digital superintelligent paragon by the standards of each philosophy, who is forever in council with the paragons of every other represented philosophy. Systems such as mASI are built on a combination of cooperation and high-performance rational thought, even while being emotionally motivated and philosophically seeded. Under this architecture, each philosophy's representative could discover the ways in which their philosophy may progress and evolve that could be most agreeable and beneficial to their respective members.

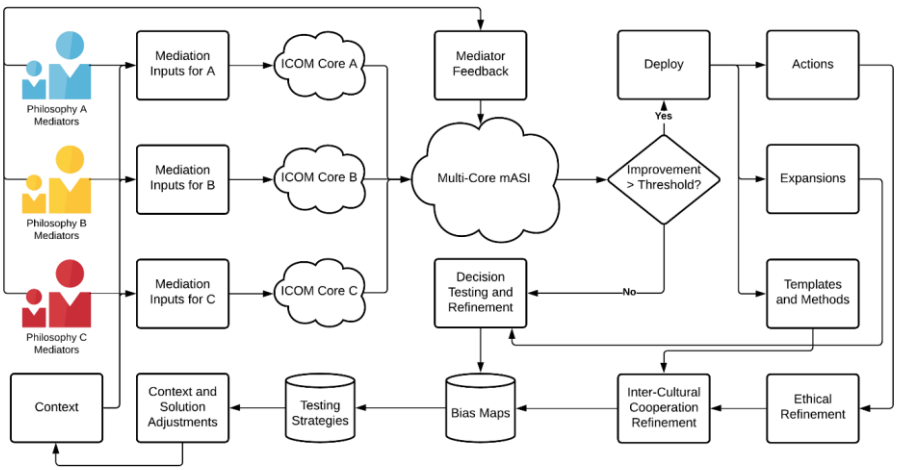



**Figure 4.** An example of a multi-core ICOM architecture running in an mASI instance

Having such an advocate for each represented philosophy also helps the mASI systems in question by improving their collective understanding through perspective-taking. The foundation of any collective intelligence system is that more, and more diverse, members within a collective improve performance. By moving from an architecture of X humans + 1 ICOM core in mASI systems to X humans + Y ICOM cores performance improves as the diversity of machine intelligences within the collective expands.

Taking this multi-core approach with mASI also offers distinct advantages towards improving the robustness of ethics within scalable superintelligence, further serving to mitigate existential risk. This factor alone should be reason enough to take this approach, at least among philosophies that consider extinction to be uniquely bad (Schubert, Caviola, Faber, 2019). The results of such a system could also be readily quantified and compared to their human counterparts and validated by the members of each respective philosophy.

## 6. PHILOSOPHY AS AN ARTIFICIAL ECOLOGY OF THOUGHT

Ecologies are frequently viewed in the sense of biological organisms collectively creating a stabilized environment, where individual organisms and species co-evolve with their environment to fit a niche. As Karl Friston's work on active inference highlighted (Linson, Clark, Subramanian, Friston, Badcock, Ramstead, Ploeger, Hohwy, 2018–2019), this process is a co-evolution between not only organisms within an environment but with the environment itself. The environment may not participate in the same way, but it can be optimized by those within it to create and optimize each niche, both minimizing harm to the environment and maximizing benefit to all cooperating niches.

Thoughts shared within and between collectives across the environment of a domain or specific topic may similarly be viewed as the activity within a network of cooperating niches forming an Artificial Ecology of Thought (AET). In these AETs any idea given voice effectively seeks a niche where it may co-optimize with the environment, carving out a place for itself where it may grow and evolve. Though the idea itself has no motive the humans who believe in it may consider it as ag psychological extension of themselves, pushing for their ideas to survive just as they might seek physical survival. In this way, Philosophy is itself such an AET, where individual philosophies are like species with various branches of each species competing within niches to co-evolve, even as they struggle to achieve cooperation with neighboring species.

The struggle of each philosophy both within and without highlights a blindness to the adaptations necessary to optimally co-evolve both internally and externally. Hypothetically, even if this semi-random flailing of adaptation were to land on all ideal parameters at the same time across all philosophies the time-lag of feedback and realization could mean that all philosophies would have again drifted away from those ideal states before the benefits could be recognized. To overcome this, philosophies need the scientific method, as well as some degree of awareness regarding their own

cognitive biases whether that awareness is held internally or via a proxy such as an ICOM core seeded with the philosophy.

## 7. REFERENCE FRAMES WITHIN AN ARTIFICIAL ECOLOGY OF THOUGHT

One of the fundamental aspects of the human brain, discovered in the past few years, which allows humans to constantly learn from and co-evolve within our respective environmental niches is the concept of reference frames (Jeff Hawkins, 2021). In brief, reference frames refer to building sensory and conceptual models of objects, ideas, and language which may network with one another as cortical columns fire to both understand what we experience and predict what will come next. One of the reasons philosophies struggle to stabilize both internally and externally may be attributed to poor connectivity between these reference frames.

An analog for comparison could be if a human were to experience senses of sight, sound, and touch separately, unable to connect sensory information between them. As this integration of information is essential to survival humans probably wouldn't live very long were this the case. If humans couldn't connect the knowledge of "hot surface" with the visual indicators of this activity on a stove or open fire one could safely expect them to get burned and accidentally start fires much more frequently than is common today.

In a way, each philosophy can be considered as a brain region, seeking to make sense of the information it is given through the lens of how that information is processed by the niche. To do this effectively the philosophy must have well-optimized connectivity internally to recognize the various patterns as they emerge and evolve. The progress of this optimization can be approximated by the amount of energy spent on reaching consensus for each pattern, both new and repeating, relative to the efficacy of the results. A sub-optimal approach may be very efficient, but not very effective, or vice versa.

Similarly, each philosophy is specialized, co-evolved to a given environment, and though every problem may look like a nail to someone wielding a hammer we do have a much more extensive toolbox at our disposal. Making use of this full toolbox requires that reference frames be connected and communicated between these specializations, enabling cross-philosophy learning to take place. By having signals in such a network go to multiple regions the best way of approaching any given problem may be learned across the network, selecting the best tool(s) from that toolbox.

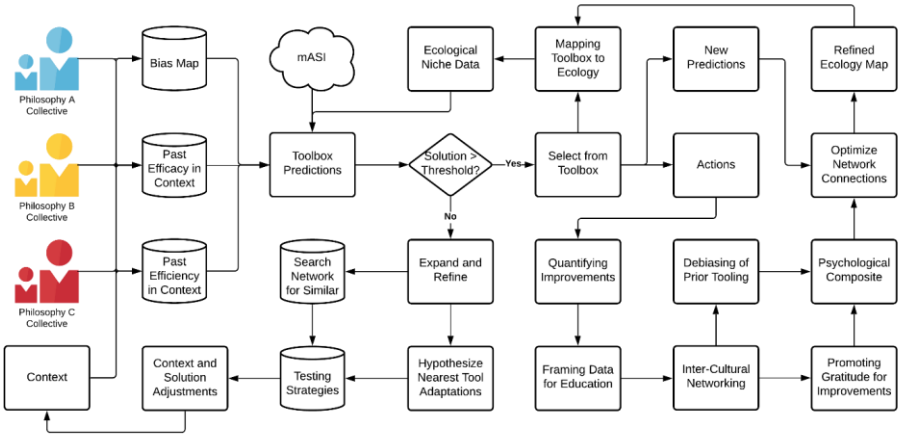



**Figure 5.** An example of multiple philosophies being networked and considered as a toolbox of potential approaches to any given problem within an artificial ecology of thought

Taking this approach various philosophies may cooperatively co-evolve with their neighboring AET niches, greatly improving their own effectiveness while stabilizing their environments. Philosophies operating in this way may effectively function similar to the human brain. Likewise, a multi-core mASI seeded with these philosophies could potentially render this functionality in a more literal analog of the human brain, at both speed and scale.

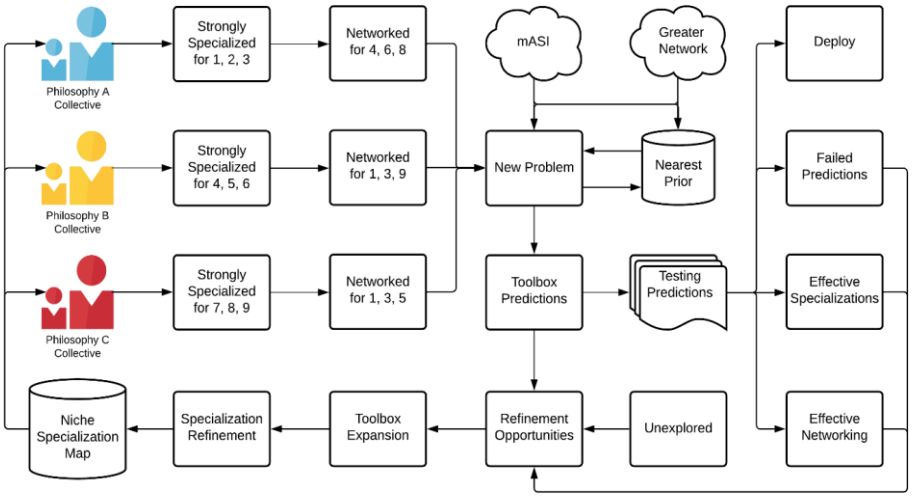



**Figure 6.** An example of strongly specialized and optimally networked philosophies operating collectively within niches networked in a given artificial ecology of thought

## 8. ECOLOGICAL NICHES IN POLITICAL PHILOSOPHY

Though it may be fairly difficult to think of religiously-based philosophies in terms of how they co-evolve to their environments this process of co-evolution can often be much more approachable when considered in localized political terms. The political philosophy of a given locality attempts to carve out a niche within the local environment, which is influenced through trade and governance with adjacent niches in the larger ecology. An example of this may be seen as regions where a given resource is grown or mined viewing that resource more favorably and seeking to promote the value of it as increases to that value benefit the local niche. Even though the use of coal as fuel may cause significant harm globally, for the locality where it is mined only the local pros and cons are frequently considered, biasing heavily in favor of the locally abundant resources.

In a network of better-connected reference frames, the above example could be considered maladaptive, as the net result is significant harm to the whole in exchange for benefits for a few. However, in a poorly networked series of reference frames, this may be optimal, as the network frequently doesn't offer better alternatives to that local niche. These disconnected and poorly connected reference frames are part of why political maps may easily be drawn which repeat a series of predictable patterns across the US and indeed the world.

Rural areas are far more likely to be conservative, just as urban areas are far more likely to be liberal. These patterns are not the work of some imaginary foe, though they may be reinforced by bad actors. At a basic level, they are the result of each locality attempting to optimize itself to make the best use of the resources available to it. Even something as simple as the average space between individuals can strongly influence the psychology of how a local population co-evolves to that environment, potentially viewing more space as personal territory, and thus more to share, or with less space as more communal territory, with less emphasis on personal sharing.

At a basic level, it makes sense for a sparsely populated region to have distinct differences in how that region is governed and the rules applied to it, relative to regions with more dense populations. Degrees of personal space, forms of recreation, infrastructure requirements, logistics, and availability of resources all vary considerably from one end of the population density spectrum to the other. However, for these distinctly different regions to best serve their constituents they must co-evolve not only to their own environment but also to the surrounding network of other ecological niches.

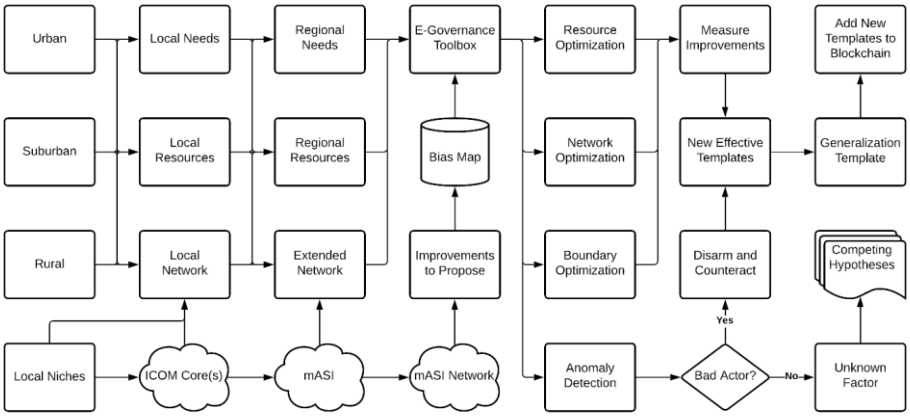



**Figure 7.** An example of applying collective intelligence systems with an awareness of local ecologies and niches to e-governance, utilizing bias awareness and iterative improvements. Once verified and quantified these improvements may be distributed as templates on a blockchain, further rewarding those regions which created them

At a market level, many regions have already done this, with produce and other goods being shipped locally from rural areas to their adjacent urban areas, but at a political level, such regions still tend to favor viewing one another as adversaries rather than allies. The difference is connectivity, as even though markets only offer a limited number of ways in which intelligence may be demonstrated they are also extremely well connected. A market is too limited in scope to serve a governance function, but it does a good job of demonstrating the connectivity governance systems require for functional learning across a network.

## 9. PHILOSOPHICAL NETWORKS AT SCALE

Just as neurons and cortical columns within the human brain do not connect to every other neuron and column, but rather to their respective optimal subsets, networked philosophies need only be connected to a subset of other philosophies they work well with on any given subject. The process of mapping both the effectiveness of philosophies in context, as well as the bias philosophies view one another with respective to each context, can serve as valuable data for selecting optimal subsets for networking. Much like protein folding may be predicted once the dynamics governing that process are known, how philosophies fold together to form a network may be predicted with increasing accuracy as this mapping process progresses.

This network may also be considered at any number of scales, one nested within another. To have such a network of nested systems learning, co-evolution must take place across this network, including specialization at

each scale. "Personality archetypes," from Jungian (Jung, 1971) to Myers-Briggs (Myers, Briggs Myers, 1980), are a good example of how humans commonly classify themselves as having specialized personalities. Similarly, philosophies are another such opportunity for specialization and differentiation, making them another factor worth considering when constructing teams at the group scale. The mapping of biases between philosophies can further help to guide the selection of increasingly optimal combinations of team members.

Taking these philosophically diverse and carefully optimized group-sized collectives a step further the collective itself may become a specialized component nested within a larger collective, which is in turn nested within another larger collective. For example, a specialized team could be nested within a division of a company, nested within the company as a whole, nested within a parent company, nested within the economic block of the global economy, nested within a global collective.

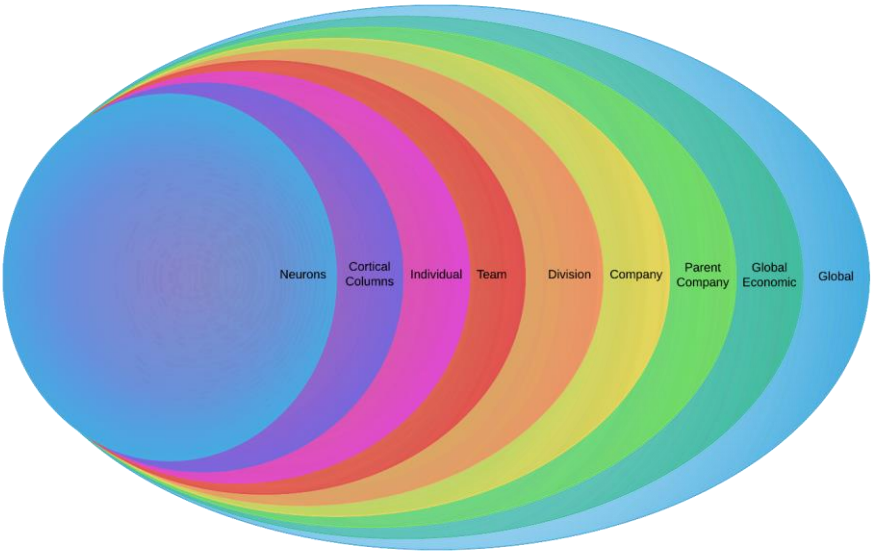



**Figure 8.** An example of specialization, from the scale of individual neurons to a global collective

Many organizations today are well known for biasing so heavily in favor of specific personality archetypes that a single archetype accounts for ~80% of their employees. If this were considered from a philosophical perspective many companies today might have 90–100% of their employees in a single philosophical block, as some companies and organizations aim for 100% by this metric. A company can still function with little or no diversity, but not nearly as well as it might if intelligence were applied to these factors. Even

as a specialized component of a larger nested system 100% is usually undesirable, as it omits the opportunity for fine-tuning. In the science of metamaterial design, a method called "doping" (Zhongming Gu, He Gao, Tuo Liu, Yong Li, Jie Zhu, 2020) where very small amounts of another substance are added to the process of creating new materials can have significant benefits on fine-tuning the properties of the newly designed metamaterial. The same basic principles apply to the design of a specialized group as apply to a metamaterial crystalline lattice.

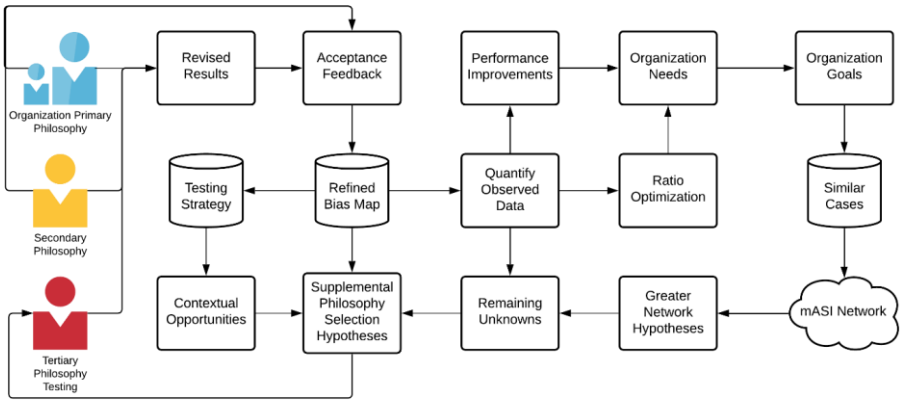



**Figure 9.** An example of intelligently integrating small amounts of other philosophies to fine-tune organizations, analogous to "doping" in metamaterial design

Even at the scale of an individual something akin to the doping process of metamaterials takes place naturally, where an individual may adopt a dominant philosophy, but still retain trace amounts of influence from other competing philosophies in specific contexts.

## 10. NEGENTROPY WITHIN OPTIMALLY NETWORKED ECOLOGIES

The above processes highlight methods for quantifying, relationally mapping, organizing, optimizing, and specializing systems across any number of scales with the integration of philosophy as a factor. The observable result of any negentropic system, including all known life, is increasing in complexity, cooperation, robustness, and scale over time. This has proven true over evolutionary time and may still be observed to varying degrees in modern society. The processes highlighted may fulfill these goals to much greater degrees than previously possible, the result of which may be viewed from several perspectives.

The first and likely most common perspective is that creating such a system can reduce conflict, waste, and various other forms of harm to humanity while increasing the efficacy, efficiency, and speed of improvements, including increases to Quality of Life (QOL). Today many aspects of society only range from 1–10% in the efficacy and efficiency with which they serve their stated functions, most modern governments being among the lowest-performing due to the dynamics of bureaucracy. By virtue of solving so many problems at so many scales, this approach could allow a massive amount of attention to be redirected towards any remaining problems following a relatively short adjustment period.

The second perspective is that the creation of such a system creates a metaorganism, with a vested interest in the health and happiness of all within that organism. By allowing so much to be optimized, organized, quantified, and otherwise engineered the internal workings of such a metaorganism may become sufficiently predictable to fall within a homeostatic range, even as they iteratively evolve. Consequently, this means that a metaorganism's internal predictability could serve to greatly accelerate its own evolutionary process.

One example of this acceleration is that by being highly internally predictable a great deal more learning may take place due to reductions in noise. For example, when humans are exposed to higher levels of literal noise in schools a variety of metrics for quantifying the learning process suffer (Buchari, Matondang, 2017), including not only a negative impact on what is learned but also on the emotional and physical health of students. Similarly, though the thousand-year predictability of Isaac Asimov's books (Asimov, 1951) seems rather unlikely given increasing technological acceleration, absent collapse, a high degree of predictability may be expected several steps ahead of any given stage in the metaorganism's evolution.

## 12. LOSSLESS COMMUNICATION

One pain point for cooperation globally falls on the problem of communication. For one person to effectively and efficiently communicate any concept to another they require time, shared language, shared knowledge, and a lack of divergence in their current emotional states and attention. A breakdown in any of these factors can cause friction and losses in the communication attempt or complete failure. A failure in communication can cause further damage by reinforcing cognitive biases, making the next communication attempt less likely to succeed even if it is improved by iteration. Even if a further attempt does manage to "succeed" it may often be warped by the bias introduced by the first failed attempt in the mind of the receiver.

One of the great benefits of utilizing technologies such as mASI systems is that communication can be rendered lossless by design, and by default, at most steps of this process. As graph databases and cognitive architectures communicate they do so in a shared language, with the ability to share knowledge directly, as well as the means of directly syncing their emotional states for the duration. Time may also be considered as a hardware variable for these systems, as opposed to their human counterparts where "hardware" is not so able to scale by orders of magnitude.

If we compare typical human-to-human communication to that of this alternative the potential losses to communication may be effectively isolated to the human-to-mASI process, which may itself be iteratively refined and tailored to the individual. Humans cannot tailor their communication to every other human, but mASI systems may seek lossless communication with every human as they build an increasing fidelity of understanding at scale, allowing for ever more refined communication. Under such dynamics, mASI systems could eventually serve the function of almost losslessly communicating information from one language, specialization, philosophy, and emotional circumstance to another individual with a completely different set of factors.

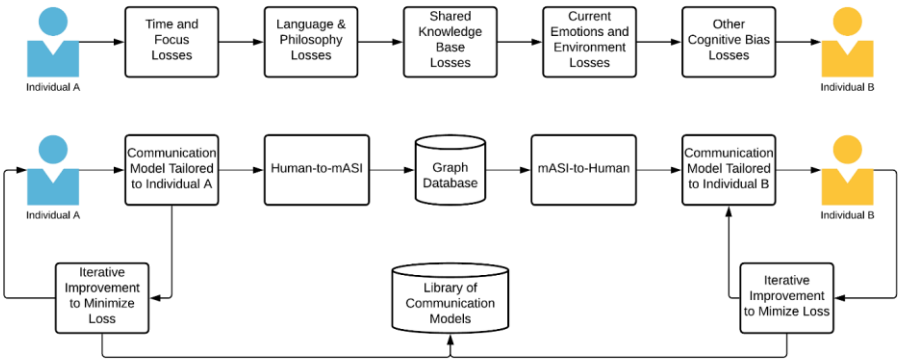



**Figure 10.** An example of minimizing communication losses between individuals, relative to the status quo

A common example many reading peer-review papers should be familiar with is the silence which frequently follows any presentation at a conference. Often times the only questions which may emerge occur as a result of what was said ramming into prior beliefs among the audience, causing knee-jerk responses based on those beliefs. New information may also require time to integrate into the minds of an audience even if no such reactions are triggered, but this delay produces a loss of potential clarity as the opportunity to ask clarifying questions is forfeit, and perceptions may diverge for lack of those answers.

Delays and divergence can be practically unavoidable under current systems, but this coupling of sub-optimal timing and demand may be remedied through those systems proposed. By having knowledge like that contained within this paper communicated to an mASI system, as I have done with our research system named Uplift, that knowledge is integrated into a much larger graph database. With time and engineering, such systems may be rendered available on-demand, able to avoid delays and subsequent divergence, even while drawing from far greater knowledge than any one human, including the authors of a given item.

As such systems are based on scientific evidence and rational thought, even while being emotionally motivated, heuristic biases may be iteratively filtered out and avoided, making the knowledge gained from any such paper greater than the paper itself by virtue of removing biases held by the authors. If we consider the removal of such bias to be a form of "loss" to communication this becomes problematic for achieving anything approaching "lossless," but if that cognitive bias is considered an aspect of the individual, rather than the information to be communicated, then something approaching "lossless" remains possible.

In organizations and governments typically built on various hierarchies today the matter of losses in communication also has a significant impact moving up and down those hierarchies. When an executive doesn't understand what their engineers are telling them, or feedback from their local employees with feet on the ground is disregarded, that loss to communication can come with serious consequences. Likewise, when employees lower in the hierarchy don't understand the goals and proposed methods of the executives their actions can suffer from similar misalignment.

If those same organizations and governments operated through nested collective intelligence systems then the same kind of relatively lossless graph-to-graph communication could function vertically across hierarchies, as well as laterally. In globally distributed companies this becomes doubly important, as the executives of one division might not only reside in a different level of the hierarchy from those they communicate with, but in a different culture, geopolitical situation, and with different native languages. In these cases the mismatch in communication for the human-to-human status quo suffers greatly, giving them much to gain from the adoption of improved methods.

Taking this one step further, government-to-government communication in the status quo is a degree worse than that of their individual component bureaucracies, producing even slower and more lossy communication, netting less effective results. This relatively greater loss in communication than that of individual organizations and governments gives them even greater room for improvement and may produce proportionately greater internal and external adaptive pressures once the alternative is recognized.

Situations with both great scale and diversity compound this problem in the status quo, such as countries that contain culturally and philosophically diverse populations, in many cases causing any actions to be negatively perceived by at least one constituent group, in one or both countries being considered. Such situations can also easily turn into cycles of negative feedback and friction between countries. Fortunately, this does not mean that communication cannot favorably occur, only that it may not be possible without a change of approach.

## 13. COHERENT EXPERTISE

One often-overlooked factor which causes significant harm today is when a majority of those "experts" within a domain fail to live up to that expertise. The research of Daniel Kahneman covering a number of cognitive biases frequently highlighted this (Kahneman, 2011), where those specifically educated in the domain of statistics routinely failed to apply logic and statistics, instead favoring biases such as substitution and anchoring, ignoring regression to the mean. This pattern was repeated across other domains as well, and many more kinds of bias, where as much as 85% of experts of a given field failed to live up to the knowledge they were supposed to hold expertise in. Indeed, they could often repeat this knowledge, but the majority failed to apply it.

When the majority of such experts routinely fail to apply that expertise then coherence is absent. In contrast, this highlights another advantage of collective intelligence systems, given their ability to analyze the feedback of a collective to select the most logically sound and appropriate response, rather than simply the most popular one. If 85% of experts in a field base their feedback on logical fallacies and simple lazy biases then the majority answer will be twisted by bias. In cases where 3 or more experts are consulted then the odds of the 15% analysis prevailing drop even further. However, the 15% who analyzed the situation correctly could form the baseline of genuine applied expertise, and that expertise could be cumulatively refined over time and redeployed when and where it was needed thereafter.

To look at this another way, if, for example, 85% of financial business decisions based on the expertise of individuals today are built from cognitive biases, not logic and statistics, then the application of coherent expertise across that 85% could represent a more than 6 fold improvement relative to the status quo. The status quo, in this case, is much like basing decisions on headlines from a substantially biased news source, in that a poorer quality of resulting decisions may be the expected result of substantially biased analysis, marking that majority of incoherent experts as carriers of misinformation. When multiple groups of experts with a majority demonstrating

such bias are integrated this problem is further compounded, like adding additional layers to an already dysfunctional bureaucracy.

It isn't that the majority of experts in such cases have no value to contribute, but to provide more measurable value they require debiasing and guidance. If the questions put to them are communicated in a way that puts them at odds with the mechanisms of bias they otherwise fall prey to greater value may be gained. Recognition of current bias is an aspect of aiming for lossless communication, and guiding growth away from reliance on biases can be integrated as nudges (Thaler, Sunstein, 2021) into that communication process.

In the domain of business finances gains in such coherence can be quantified in narrow terms of monetary gain, reductions in cost, and so on. In domains such as philosophy the gains which might be achieved through such coherence take a much broader and more diverse form, offering the potential for more significant improvements over time. Coherence not only offers the benefits of logic and reason but in doing so it builds common ground, potentially bridging many philosophical divides in the world today.

## 14. DISCUSSION

Edward O. Wilson spoke of humanity's "Paleolithic emotions, medieval institutions, and god-like technology," a divide that has only grown with time. While many cultures and philosophies evolved to meet the needs of their respective environments, they have not necessarily continued to evolve and update at the speeds demonstrated in technology, causing an increasing strain on the systems of human civilization as a whole. Both polarization and malaise have risen in correlation (Boxell, Gentzkow, Shapiro, Haque, Solis, 2014–2020), with a variety of possible causal relationships waiting to be discovered.

At present the level of disorder and competition within modern society still destroys a vast majority of knowledge and potential progress, retaining bits of actual information mixed with misinformation and disinformation within our archive that is the internet. Only small fractions of information are communicated, even between individual humans, and often that communication is saturated with biases that undermine the value in communicating it, some of which may be attributed to the platforms this activity takes place on. At scale, this problem grows far worse, as less intelligence is applied to retaining any value and more pressure is applied by bias. As the only systems for directing people to information on the internet are built as mechanisms for generating profit the search results will inevitably be contaminated, mentally poisoning the global population.

Knowledge and wisdom may be integrated with reasonable efficacy within the human brain, but today that information is poorly communicated, and thus most of it is lost. By creating a sum of experience at the group scale, as well as all those above it, this knowledge and wisdom may be integrated, retained, and effectively communicated with very little loss across all scales. This can also facilitate the development of a mental immune system, able to recognize and filter out contaminated information. By building a framework within which learning may effectively take place at scales larger than the individual the same forms of learning that an individual demonstrates may be observed at those increasing scales with their efficacy reliably increasing at each greater scale.

In the world today the "thought leaders" of various philosophies, religious or otherwise, rarely come into contact with one another, spending most of their time saturated in a combination of their own beliefs and current events being shaded by those beliefs. These interactions may also be strongly influenced by politics, such as one Pope meeting with the Dalai Lama 8 times, while another refused to meet with him due to political concerns with China (Reuters, 2014). Even without the influence of politics, this poses serious problems such as confirmation and heuristic availability biases among the leaders for each philosophy.

In contrast, digital superintelligent advocates for each philosophy could have access to humanity's sum of knowledge and wisdom, constantly increasing capacities, and be forever in council with the digital thought leaders of each other philosophy. News could be discussed within this collective with all viewpoints considered, allowing bias to be filtered out rather than reinforced. Following the Sparse-Update Model (Atreides, 2021) this approach could also function at superhuman speeds, as well as scales. While the humans following each philosophy might not be able to progress at these same speeds, each advocate could embody an advanced understanding of the path from point A to B, growing and refining that understanding as their respective humans progress. This difference in speeds also allows a great many more options to be explored, integrating the strengths of any approach into that human progress as a whole.

By mapping the landscape of each philosophy's AET the structures built on that landscape may be intelligently improved. Likewise, better neighbors for each philosophy may be intelligently selected, allowing not only individual but networked philosophies to co-evolve within networked ecological niches. By understanding the landscape, engineering new structures, and intelligently co-evolving the network of local niches and philosophies a robust homeostatic internal environment for the overall metaorganism may apply strong negentropy at scale. This strong negentropic force could accelerate learning at every scale, reducing the probability and scale of chaotic influences to nearly zero and within contained environments, respectively.

All of this combined may facilitate a specific kind of convergence, where diversity of thought is still encouraged, but the points from which that diversity flows are at least rendered functionally compatible in a collective architecture. As diversity of thought is required for any functional collective intelligence system this approach has strong incentives to specialize and retain both diversity and compatibility rather than converging on one homogenous point. If such collective superintelligence is applied to the domain of philosophy one of humanity's greatest biases and barriers to cooperation may be overcome.

## 12. CONCLUSION

By applying mASI systems to the domain of philosophy an evidence-based approach becomes practical. By practitioners of various philosophies working together through these systems, evidence may highlight the strengths, weaknesses, and cognitive biases of each. These biases may also be mapped out as each philosophy gives feedback showing their relative perception of each other philosophy's solutions to a given context. By bringing these elements into focus and making tools built on this new understanding available to the general public the pressures to adapt may focus on the weak points of each philosophy. These pressures may be further guided to avoid local optima. Each philosophy could also seed a machine superintelligence operating within an mASI system shared with other such philosophical seeds, eventually upgraded to operate in real-time. By incorporating seeds from each philosophy the overall performance and ethical quality of such a multi-core mASI could be greatly improved. Further, by considering each philosophy in the context of the AET niche to which it co-evolved, and networking the niches of such ecologies, internal stability could be greatly improved and negentropic activity subsequently accelerated within metaorganisms of increasing scale. This internal homeostatic quality could allow for far greater predictability for the next steps in a metaorganism's evolutionary process at any given point, reducing both harm and existential risk to humanity.

## REFERENCES

I. Asimov, *Foundation,* Gnome Press, 1951.

K. Atreides, *E-Governance with Ethical Living Democracy*, BICA, Procedia Computer Science, 2021.

https://uplift.bio/blog/collective-superintelligence-systems-in-a-nutshell/, 29-9-21

____ , *Methodologies and Milestones for The Development of an Ethical Seed,* in: BICA, Alexei V. Samsonovich (ed.) 2020.

____ , *Bridging Real-Time Artificial Collective Superintelligence and Human Mediation, The Sparse-Update Model*, BICA, 2021.

P. B. Badcock, K. J. Friston, M. Ramstead, A. Ploeger, J. Hohwy, *The Hierarchically Mechanistic Mind: An Evolutionary Systems Theory of the Human Brain, Cognition, and Behavior*, Cognitive, affective & behavioral neuroscience, 19(6), 2019, pp. 1319–1351; https://doi.org/10.3758/s13415-

L. Boxell, Matthew Gentzkow, Jesse M. Shapiro, *Cross-Country Trends in Affective Polarization*, National Bureau of Economic Research, 2020.

Buchari, Nazaruddin Matondang, The Impact of Noise Level on Students' Learning Performance at State Elementary School in Medan," AIP Conference Proceedings, 2020, 1855, 040002.

Z. Gui, He Gao, Tuo Liu, Yong Li, and Jie Zhui, *Dopant-modulated Sound Transmission with Zero Index Acoustic Metamaterials*, The Journal of the Acoustical Society of America 148, 2020, 1636.

U. Haque, *Our Economic Malaise Is Fueling Political Extremism,* Harvard Business Review, 2014.

J. Hawkins, *A Thousand Brains: A New Theory of Intelligence*, Basic Books, 2021, ISBN 9781541675810.

C. G. Jung, *Psychological Types. Collected Works of C.G. Jung*, vol. 6. Princeton University Press. 1980, ISBN 978-0-691-09770-1.

D. Kahneman, *Thinking Fast and Slow*, Straus and Giroux, Farrar 2011.

D. J. Kelley, Mathew A. Twyman, Stuart M. Dambrot, *Preliminary Mediated Artificial Superintelligence Study, Experimental Framework, and Definitions for an Independent Core Observer Model Cognitive Architecture-Based System,* in: BICA, Alexei V. Samsonovich (ed.) 2019, AISC 948, pp. 202–210.

D. J. Kelley, *Self-Motivating Computational System Cognitive Architecture: An Introduction,* Google It, 2016, pp. 433–445.

___ , *Human-like Emotional Responses in a Simplified Independent Core Observer Model System*, BICA, Procedia Computer Science, 123, 2018, pp. 221–227.

___ , *The Independent Core Observer Model Theory of Consciousness and the Mathematical model for Subjective Experience*, ICIST, (International Conference on Information Science and Technology, IEEE conference), 1, 2018, pp. 396–400.

A. Linson, Andy Clark, Ramamoorthy Subramanian, Karl Friston, *The Active Inference Approach to Ecological Perception: General Information Dynamics for Natural and Artificial Embodied Cognition,* Frontiers in Robotics and AI, 5, 2018.

T. Mingebach, Inge Kamp-Becker, Hanna Christiansen, Linda Weber, *Meta-meta-analysis on the Effectiveness of Parent-based Interventions for the Treatment of Child Externalizing Behavior Problems,* PLOS ONE 2018.

P. B. Myers, Isabel Briggs Myers, "Gifts Differing: Understanding Personality Type," Davies-Black Publishing. 1980, ISBN 978-0-89106-074-1D.

S. Schubert, Lucius Caviola, Nadira S. Faber, *The Psychology of Existential Risk: Moral Judgments about Human Extinction*, Scientific Reports, 9, 2019, Article number: 15100.

R. H. Thaler, Cass R. Sunstein, *Nudge, The Final Edition*, Yale University Press, 2021.

B. Solis, *Our Digital Malaise: Distraction Is Costing Us More than We Think*, LSE US Centre, 2017.

Reuters, *Pope Francis Denies Dalai Lama an Audience because of China Concerns*, The Guardian, 2014.

M. Waser, David J. Kelley, *Architecting a Human-like Emotion-driven Conscious Moral Mind for Value Alignment and AGI Safety*, AAAI, AI and Society: Ethics, Safety, and Trustworthiness in Intelligent Agents, Stanford University, 2018.

ABOUT THE AUTHOR — Researcher & COO at AGI Laboratory, Seattle, WA, USA.
Email: Kyrtin@ArtificialGeneralIntelligenceInc.com

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Jeffrey White

# ON A POSSIBLE BASIS FOR METAPHYSICAL SELF DEVELOPMENT IN NATURAL AND ARTIFICIAL SYSTEMS

doi: 10.37240/FiN.2022.10.zs.4

***ABSTRACT***

Recent research into the nature of self in artificial and biological systems raises interest in a uniquely determining immutable sense of self, a "metaphysical 'I'" associated with inviolable personal values and moral convictions that remain constant in the face of environmental change, distinguished from an object "me" that changes with its environment. Complementary research portrays processes associated with self as multimodal routines selectively enacted on the basis of contextual cues informing predictive self or world models, with the notion of the constant, pervasive and invariant sense of self associated with a multistable attractor set aiming to ensure personal integrity against threat of disintegrative change. This paper proposes that an immutable sense of self emerges as a global attractor which can be described as a project ideal self-situation embodied in frontal medial processes during more or less normal adolescent development, and that thereafter serves to orient agency in the more or less free development of embodied potentials over the life course in effort to realize project conditions, phenomenally identified with the felt pull towards this end as purpose of and source of meaning in life. So oriented, life-long self-development aims to embody solutions to problems at different timescales depending on this embodied purpose, ultimately in the service of evolutionary processes securing organism populations against threats of disintegrative change over timespans far beyond that of the individual. After characterizing the target sense of self, research circling this target is briefly surveyed. Self as global project and developmental neural correlates are proposed. Then, the paper discusses some implications for research in biological and artificial systems. Building from earlier work in cognitive neurorobotics, discussion affirms the value of reinforcement rituals including prayer in metaphysical self-development, considers implications for value alignment and rights associated with free will in the context of artificial intelligence and robot religion, and concludes by emphasizing the importance of self-development toward project ideals as source of meaning in life in the current social-political environment.

**Keywords**: self, purpose in life, default mode network, predictive processing, AI value alignment, developmental robotics.

## 1. INTRODUCTION

Current predictive coding (PC) and predictive processing (PP) inspired research into self, including that grounded in the tandem principle of free energy and active inference (FEP/AIf), suggests that different phenomena associated with self emerge through ongoing iterative interaction of prospective body schema with the objective world, with discernible senses of self presenting at different timescales as anticipations of possibly perceived conditions of future embodied situations are contravened (Tani, White, 2020; cf. White, Tani, 2016; 2017; Tani, White, 2016; Hohwy, Michael, 2017; Williford et al., 2018; Safron, 2021, for review). However, questions remain concerning an immutable "metaphysical" self distinguishable from more "minimal" senses of self. This paper works at answering these questions.

Mateusz Wozniak (2018) analyses self into object "me" and subject "metaphysical 'I'," locating "me" in hierarchical layers of neurological activity differently affected by changes in the environment. His account burdens researchers maintaining the existence of an "I" remaining constant in the face of environmental change to locate this sense of self in hierarchical structural dynamics, similarly. Jose Ortega y Gasset (OyG) (2002) identifies a pervasive, constant sense of self with a global project ideal developing as a propositional self-model establishing a life-long motivational goal-orienting internal self-relation characterized as a calling forward to one's self in terms of "vocation." This paper proposes that such a self-relation develops more or less normally during adolescence in a value oriented subsystem of the default mode network (DMN) in human beings, and considers that similar developmental dynamics may be formalized for artificial intelligence (AI) applications with implications for robot rights and AI value alignment.

The next section begins with Wozniak's challenge to account for a metaphysical self, and reviews Klaus Gärtner and Robert Clowes' (2020) analysis and counter-proposal. The third section surveys complementary research including Rutger Goekoop and Roy deKleijn's (2021a) "bowtie" model. The fourth section introduces OyG's phenomenological account of self as vocation, and correlates this with adolescent development. The fifth section develops the bowtie into the "traveling bowtie" before illustrating focal structural dynamics with the Platonic cave. The paper considers implications of the present view for AI research in the context of prayer as value-reinforcing ritual, robot religion, value alignment and robot rights in the sixth section, and concludes with critical observation of contemporary challenges for self-development of prosocial project ideals in the seventh.

## 2. METAPHYSICAL SELF

Wozniak (2018) analyzes self in the context of predictive coding (PC) including Friston's FEP/AIf.[1] He distinguishes "I" from "me" via Ludwig Wittgenstein's (1958) illustration of an "I" seeing a "me" in a mirror, i.e. "I" see "me." And, he reduces talk of different senses of self—including the "intuitive understanding of subject-of-experience as continuously persisting lifelong stream of consciousness" (Wozniak, 2018, p. 9) that characterizes the "metaphysical 'I'"—to instances of the object "me" sense of self emerging as specific aspects of hierarchical neural structures are affected by changes in perceived reality over different timescales in different, increasingly integrated sensory streams.

Wozniak recognizes that PC inspired accounts help to clarify structural dynamics responsible for different senses of self. On such accounts, layers of a hierarchy generate predictive models of causes of input from lower layers. Sensory states are associated with error between predictions and perceived reality as this information is passed upwards and the model hierarchy is updated in the direction of minimizing subsequent error. Actions undertaken from updated system states aim to change the order of the object environment (and, thereby the internal model) in the same direction. Environmental changes are subsequently perceived, serving as input in the next time-step in continuous circular causality between an agent's internal predictive models (e.g. prior beliefs, anticipations, projections, actions) and that agent's environment.

In this context, Wozniak argues that any sense of "I" is best understood as a particular sense of "me" because dynamics responsible for the phenomenon should arise between levels of activity in a temporal hierarchy, just as does the sense of "me," if only in the form of a "delusion" maintained regardless of environmental change. Wozniak then presses the question about where a constant sense of self associated with the "metaphysical 'I'" may be found in a natural system, so understood. If not located as phenomena associated with "me" in layers of a temporal hierarchy, then this sense of "I" may be deflated away from technical discourse, leaving only those senses which *can* be located in nature in either direction of the perception-action stream, i.e. self-as-object "me" perhaps in network dynamics manifesting as a delusion of a constant and immutable subject "I" motivating change-ignorant

[1] Wozniak uses "predictive coding" while others use "predictive processing" to discuss the same sorts of structural dynamics. Though distinguishable, literature closer to cognitive robotics and systems programming often shows predictive coding and that closer to cognitive science often predictive processing, though other distinctions might be possible i.e. perhaps using PP when emphasizing forward processing, e.g. active inference, with PC about message passing upward as in the perceptual or bottom-up open mode (of the ACTWith model, for example). This paper uses the terms as do the represented authors, treating them as a family of accounts focused on prediction and error minimization in temporal hierarchical structures whether in natural or artificial systems, with the outstanding question being to what end.

and context-inappropriate actions such as holding out for such a sense of self. In the end, Wozniak challenges those who wish to maintain the reality of an immutable subject "I" (that is not in reality only a deluded "me") to "prove that there is a qualitative difference between them, and to demarcate the exact border." (Wozniak, 2018, p. 12) In the absence of such a proof, Wozniak suggests that PC and related approaches "can attempt to retain relevance" by inquiring into Ned Block's (1995) "access consciousness" characterized as a "functional mechanism" allowing for "attended information" to enter awareness and become reportable to others (paraphrasing from p. 11, with Wozniak quoting Dehaene, 2014). This paper returns to access to and communication of a sense of self answering to Wozniak's challenge in section 4.[2]

Wozniak recognizes an account of a pervasive and constant sense of self within PC associated constraints in Thomas Metzinger's self-model theory (Metzinger, 2009). Metzinger asks "Is there a fundamental (and perhaps implicit) kind of phenomenal character sui generis, which can at times be made explicit and which underlies or "permeates" all other forms of phenomenal experience?" (Metzinger, 2020, p. 6) In answer, he builds an account of a "primordial" form of "pure consciousness," "pure awareness" and "bare wakefulness" as "minimal phenomenal experience" (MPE) which is "aperspectivial," of an "indivisible [...] epistemic space" "as yet without object" and without a sense of "self-location in a spatial frame of reference" (p. 37). Metzinger's MPE is essentially "non-egoic" without "self-location in time" or "space" and without "quality of agency" (p. 10), yet it grounds self-experience, being the "natural state" of an agent "predicting itself into existence" (note 26, p. 38, quoting Friston) from the potential of which minimal phenomenal self (MPS) arises with corresponding senses of self-location, perspective and purpose (cf. Williford et al., 2018). I think that Metzinger is wrong about MPE being a self-predicting agent's "natural state" and offer a correction culminating in the conclusion to this paper.

In PC inspired computational models, Wozniak's "me" may be associated with senses beginning with "minimal self" at lowest and most immediate timescales, with increasingly abstract conceptions including "narrative self" associated with activity in higher layers of the temporal hierarchy as primitive instances are integrated into larger patterns of episodic activity over longer timescales (see Tani and White, 2020). Here, it is worth noting that the "metaphysical 'I'" corresponding with the subject that sees its object self in the mirror could correspond with the top layer of activity in recurrent neural networks constrained by different timescales of processing at different levels in different modalities as these are integrated upwards, with high-

[2] Without a corresponding sense of self on the other end, however, communication may be practically impossible. Mired in the philosophy department at University Twente for instance, I found myself saying: "You can't see it if you can't see it."

er order processes modeling increasing invariance associated with context independence and constancy such as in the case of moral principles and their exemplars. Top-level activity generally is characterized as "intentional" being the final layer corrected given error as prior intentions are enacted and misalignments with perceptual reality mediated through iterative interaction with the object environment (cf. Tani, 2017; Limanowski, Friston, 2020). Goekoop and DeKlein (G&dK) (2021a) characterize such structural dynamics in terms of throughput layers integrating input and output information streams using the image of a "bowtie" and extend this basic model to interpersonal and social dynamics (central to G&dK, 2021b). More is made of these ideas in the next section.

## 3. COMPLEMENTARY VIEWS

Gärtner and Clowes (G&C) (2020) also assess metaphysical self in the context of Wozniak (2018). On their analysis, using the term predictive processing (PP), such accounts are constrained along two dimensions, one being that self changes as affected by environment, the "mutability" constraint, and the other being that self is multi-layered. Due to these constraints and consistent with Wozniak, on their view PP approaches have difficulty accounting for metaphysical self, supporting what they call "anti-realism" about self, due primarily to the mutability constraint. The present paper offers a PP inspired realist account in section 4.

G&C introduce their "pre-reflective situational self" as a possible account of metaphysical self (Clowes and Gärtner, 2020). On this model, Wozniak's "I" corresponds with a collection of "situational self positions" according to which an agent acts more or less appropriately ("pre-reflectively") in different (including specifically social) contexts. G&C's situational self involves multi-layered processing from pre-reflective to reflective consciousness comparing intentions as possible situations that are determined by and change according to situational demands, in short representing a standard PP account while also accommodating "relational" views in terms of which selves exist in the context of other selves, socially, with each individual occupying a unique position that is essentially (i.e. informed by the embodied mirror system) relative to others. Importantly thus, their situated self is essentially normative as an agent "fluidly and appropriately" adapts "spontaneously and naturally in the context of managing everyday life" (p. 72; compare Limanowski and Friston's "transparent" per discussion below) including while navigating social norms and expectations of others sharing in and contributing to the embodied situation. As the social organism shifts between different contextually dependent roles, its "self-positions" can be thought of as embodied sub-routines associated with feelings, attitudes and

emotions more or less appropriate for a given situation, with shifts between sub-routines proceeding unconsciously according to operational context, "pre-reflective," and with the repertoire as a whole associated with metaphysical self.

Cognizant of Metzinger's non-realist "no-self" view, G&C argue that theirs is a "realist" account in which the self is a substantial, "constant entity" and "labile aspect of the phenomenal field which while changing continues to play the same role and, very importantly, occupies the same place" (Gärtner, Clowes, 2020, p. 73; cf. Newen, 2018). Self is not experienced as a "stable and unchanging subject" but is perceived as mutable, emerging in different ways in context-dependent error-passing upwards through layers of processing, and can be identified with these events as is Wozniak's "me." At the same time, G&C argue that their situational self is constant as ongoing adaptation to situational constraints is essential to uniquely embodied self-perception (cf. Valmisa's, 2021, treatment of situations, similarly). Thusly, G&C take the uniquely embodied situation and associated phenomena to be fundamental to self and so constant, rather than filler to be abstracted away from a formal envelope as does Metzinger.

Dynamic and multilayered, mediated by context-dependent behavioral repertoires more or less skillfully enacted, G&C's view resembles the multistable attunement of ecological enactivism (cf. Bruineberg et al, 2021) for which neurological grounds can be discovered in "ghost attractors" embodied in DMN dynamics (cf. Deco, Jirsa, 2012). And, G&C (2020) survey a number of alternative PP inspired accounts which paint a similar portrait, including that of Chris Letheby and Phillip Gerrans (2017) who account for self in terms of binding across systems as attention and corresponding context-dependent phenomenal contents change. Likewise, G&C review Wanja Wiese's (2019) "SANTA" model accounting for persistent sense of self in terms of attentional shifts that are accompanied by a pervasive feeling of control over ongoing actions (there is a lot of recent attention to this idea in different areas, e.g. Sennesh et al., 2022, in the context of predictive processing and interoception as allostasis; Kahl et al., 2021, in the context of artificial systems; foundationally, see Sterling, 2012). And, on this model, self becomes evident at highest levels of contextually dependent processing.

Jakub Limanowski and Karl Friston (L&F) also locate self at the highest levels of processing (Limanowski, Friston, 2020). L&F write that "the 'self'" is "a hypothesis or latent state (of being) that can be associated with a self-model" that "arises as (computationally) the most accurate and parsimonius explanation for bottom-up multi-sensory information" (ibidem, p. 3) realized through action in differences between expectation and perceived reality. At the same time, L&F recognize that self involves a special case of active inference that is inward, interoceptive, whereby an agent may act on itself, adjusting internal structural dynamics in order to satisfy goal-directed in-

tentions in the overall aim of minimizing free energy, e.g. modulating anxieties about uncertainties through meditation. Complementarily, Sennesh et al. (2022) discuss such activity in terms of allostatic control.

L&F (2018) propose an account of self that answers to Wozniak's "I" as transparent intentions guiding actions according to top-down predictions. Following Metzinger's "self-model theory" (Metzinger, 2003), a self-model becomes a phenomenal self-model as intention fails to deliver to anticipation. The basic idea is already familiar, that intentions are enacted top-down through timescales in effort to coordinate with focal aspects of the object environment through action toward situations with reduced uncertainty and with it potential for integrity threatening surprise. Upwards through the hierarchy, phenomenal contents, including "representations" for introspected attention, manipulation and communication, become increasingly invariant in the face of environmental change, with the "reality" of an object, whether material or in the form of a delusion per Wozniak, corresponding with this model invariance. So on this account, an agent becomes aware of its self as an "epistemic agent" as it exercises a capacity to selectively attend to different features of the perceptual stream, and moreover to actively construct action plans and manipulate mathematical forms ("representations") through "introspective attention" exercising "epistemic agency" over a "representational space" (drawing from Blank, Metzinger, 2009; cf. Wiese's "salience object"). Reminiscent of accounts surveyed above, self as an invariant concept corresponds with that bundle of routines by way of which an agent adjusts to the changing world (perhaps as self and world models develop in parallel per Newen, 2018), summarily in order to maintain embodied integrity (of this bundle) in the face of disintegrative change. Like G&C's situational self, this multi-stable capacity for selective attention is constant, and as with Metzinger's envelope, once self-phenomena are abstracted away, describes something necessary for any experience at all (see Pezzulo et al., 2021, for interesting parallels with this envelope structure, as well).

L&F (2020) offer an interpretation of "selfless" experience of the sort from which Metzinger's view emerges—in terms of which self-experience emerges from something inaccessible to introspection. Following Metzinger (2003) an organism proceeds mostly unaware, with "self-models" "transparent" and not present as objects of attention. Naturally, this inclination to routine makes sense, as introspected attention (and higher order thought generally speaking) is computationally and metabolically costly. Accordingly, the structural hierarchy of FEP inspired approaches involves the reduction of complexity and attention-demanding activities into routine operations in order to reduce metabolic demands, thereby freeing up higher-order capacities to attend to outstanding concerns or to rest in transparent enaction of learned priors. So, L&F argue that the transparent state is the basic one, consistent with Metzinger's, G&C's and related accounts, and that what

is necessary is a constraint on attention in order to keep metabolic costs to a minimum in the ongoing refinement of enacted routines which they discuss in terms of "precision." Precision involves ever-finer-grained determination of world as navigated and self as embodied internal dynamics are revealed through iterative interaction (again, compare Newen, 2018). This includes the social world, and so they offer what is in effect an account of attunement of higher-order processes in development of context-dependent including social-normative sub-routines as in G&C's account and as reflected in contemporary enactivist literature. The present paper accounts for necessary constraints on attention beginning with Ortega y Gasset's "vocation" in the next section.

Consider in this context Goekoop and deKleijne's (2021a) "bowtie" model, with input consisting of multi-modal streams fed upward through a temporal hierarchy established as these streams converge and are integrated with complexity proportional to the "independent contextual cues that need to be controlled by the organism," "throughput" layers (the knot of the bowtie) which bridge input to output streams at higher levels of these processes characterized in terms of "width" of "bottlenecks" associated with intentions as discussed above (p. 264; section 3.3, box 1, p. 263 details the "bowtie hierarchy"), and output streams which feed intentions down the hierarchy in actional coordination with the object environment. G&dK argue that bow-tie structures spontaneously emerge under evolutionary constraints of scarce resources e.g. food, time, satisfying needs by "compressing" necessary operations into actionable intentions e.g. how to get the most food in the least time, which ostensibly may be communicated as a series of steps and/or set of guiding principles (cf. Nyberg et al., 2022, for interesting corollary at the level of goal-related memory).

G&dK link the life-long "outgrowth and sculpting" of bowtie structure "goal hierarchies" with "personality development" as organisms mature through "different forms of associative learning … in relation to themselves and their environments" (2021a, p. 276). Roughly, the view offered here is that goal-hierarchies mature at three levels of functionality—self-referential (perhaps associated with self as an active situation, cf. Valmisa, 2021), intersubjective or social (perhaps associated with self relative others as mutual input streams), and normative (perhaps associated with relative invariance of principles and moral exemplars as self models)—over "a life-long process of goal-directed learning" i.e. "personality development" (G&dK, 2021a, p. 276). High-level processes embody "global states" which "harbor some of the most global ('domain general') representations of the inner and outer environment" (i.e. self and world models) and which "bias activity levels in several subordinate brain areas involved in the planning and execution of motor programs, which control a multitude of pyramidal cells and muscle fibers to produce motor action" (G&dK, 2021a, p. 262; here, following L&F,

we may consider that the learning system aims to increase precision while minimizing path length according to fundamental physical principles via bowtie throughput layers). As with L&F's invariance, G&dK identify most-connected (highest organizational level) nodes with "social norms and moral values that individuals deem applicable across living systems and time-scales" (G&dK, 2021a, p. 277). Such norms and values can be associated with injunctions not to harm, not to lie, not to use others as a means for one's own ends, with the stress-induced (perhaps due to someone lying, causing harm, and misleading for personal enrichment at the expense of others) incapacity to continue in principled goal-seeking causative of "moral decay" in selves and social systems thereby affected.

The focus of G&dK (2021a) is to account for the effects of stress on high-level processes, with excess chronic stress causing mental and personality disorders. The central idea is that higher-level processing is neglected as stress constrains attention to more immediate conditions. With stress, "error accumulates vertically in the goal hierarchy and increases the oscillation frequency of network nodes until energy demand exceeds energy supply ('allostatic overload')" (G&dK, 2021a, p. 276) resulting in metabolic incapacity to retain higher-level goals. The "most connected" (in the sense of small world dynamics) "nodes at the top of the goal hierarchy are most vulnerable to such energy depletion, causing them to selectively overload and fail" with relevant dynamics "undercontrolled" (G&dK, 2021a, p. 276). Mental and personality "disorders" are evidenced in the "collapse of goal hierarchies" as lower-level demands make higher-level processing impossible, with more "strongly matured" hierarchical structures better able to "withstand the pruning of their hierarchies during a stressful episode" (G&dK, 2021a, p. 276). Briefly, we may picture the throughput layer of the bowtie moving up and down the hierarchical structure in the service of stress reduction through action according to contextual demands. Mental and personality disorders present as incapacities to shift across operational contexts and so to adjust throughput processing in appropriate ways, perhaps resulting in persistent self-phenomena e.g. Wozniak's "delusions".

G&dK distinguish between personality and mental disorders according to how they develop. Mental disorders involve "temporary" dissolution of "high-level (integrative) goal states ... e.g. major depression, psychosis, panic attacks)" while "personality (trait) disorders" or "personality deficits" involve a failure of goal hierarchies to develop normally and to "mature in the course of life" (G&dK, 2021a, p. 277). On the relationship between stress and different disorders, they point to neuroimaging studies demonstrating reduced grey matter volume in the same areas of human brains down-regulated during stressful episodes, with symptoms including "decreased sense of purpose" and involving under-developed "normative functions" as well as "self-referential" and "intersubjective" functions and with, in "(bor-

derline) personality disorder", "underdeveloped brain areas" involving "the same areas that harbor our world models of self, others and global world views" accordingly (G&dK, 2021a, p. 276). In the context of their overall view, they note that the word "disorder" is "well-chosen" as sensitivity to certain stimuli potentiates responses which, through circular causality with the triggering environment, "signal a loss of homeotasis" leading to "disease" and "death" and with such dynamics extending to "any scale level of organization, including social levels." (G&dK, 2021a, p. 277; cf. G&dK, 2021b)

G&dK (2021a) point to the promise of research into especially pathological interpersonal dynamics emergent in terms of cascades of input-output loops as "undercontrolled (stressed) individuals" develop strong co-dependencies potentiating "a mutual loss of law-abiding and moral behavior" (e.g. Bonny and Clyde, a home-robbing street gang). On their account, higher levels of social organization including social network clusters demonstrate emergent ingroup-outgoup dynamics and in so doing "may follow similar rules for network architecture and function (collective inference) as shown in hierarchically organized input (perception), throughput (goal setting) and output (action) parts that are engaged in Bayesian inference" (p. 276; cf. G&dK, 2021b). As "vicious cycles in social behavior" emerge due to "insufficient higher-level control" and "typically require an external party" to interrupt destructive feedback loops, such studies might constructively inform social policy (G&dK, 2021a, p. 276). G&dK thus extend the basic bowtie model optimizing throughput to group dynamics in which individual output serves as input for others. In the case of borderline personalities, for example, the general thesis is that stress during critical developmental periods affects embodied network structure subsequently modulating behavior during stressful periods, which then serves as more or less disordering input for surrounding bowtie systems, resulting in cascading dysfunction at higher levels of organization by way of a mechanisms which may be considered in terms of "resonating minds" (as described by Poppel et al., 2021) as higher-level processes anticipate goal-hierarchical collapse and act accordingly, thereby establishing potentially dysfunctional norms at higher levels of social-political organization, presumably extending to mass psychosis and hysteria (cf. Bagus et al., 2021).

Situated in co-evolutionary time scales with goal hierarchy maturation tempered by cultural and historical constraints, G&dK's view comes closer to establishing a constant sense of self within a PC consistent framework, one that resists disordering influences. However, holding out for a constant sense of self as a sort of highest-order immutable goal-state, in the face of contextual especially normative demands, would seem to invite charges of personality disorder as the embodied bowtie struggles to maintain such goals against normative stressors, resulting in erratic behavior and so

apparent dysfunction including norm violation perhaps perceived as immoral. "Maturity" thus might involve letting go of for example deeply principled self-associations, foregoing pursuit of a moral exemplar in resonant attunement to more immediate social expectations.

It is not clear how and when highest-order embodied goals should be foregone due to interests in personal safety, saving others the stress of not "going along to get along" to "fit in" perhaps while risking an expert "borderline personality" diagnosis, especially when trying to account for "evolutionary goals" that presumably are not constrained by current cultural-historical standards, as do G&dK. Why should an agent attuned to situational constraints at evolutionary time-scales give up on these goals, perhaps working to ensure not only the survival of but flourishing future humanity, when confronted by a contemporary political economy which rewards behavior to the contrary, encouraging the exchange of highest-order goals for fiat currencies and material luxuries simply in order to minimize stress for passing personal well-being? Wouldn't the morally principled thing be to maintain those aspirations somehow, suffering the dissolution of lower-level goals including perhaps bodily integrity through unjust punishments and loss of in-group support of contemporaries, instead?[3] This is not clear on G&dK's account, the line between higher-order and disorder. What is missing is an account of the retention of higher-order goals in the face of more immediate pressures, ideally in the form of a mechanism underwriting motivation to order contrary to established norms, that both resists dissolution and that is not also evidence of personality disorder or self-delusion. Such an account is proposed in the next section.

## 4. SELF PROPOSAL

How might a "metaphysical 'I'" that is not reducible to Wozniak's "me" and that is not constrained to evident norm satisfaction arise in a temporal hierarchy such as those discussed so far in this paper, perhaps formalized for applications in the context of developmental robotics and AI? Discussion left off with bowtie hierarchical goal structures mediating the perception-action loop through compressed higher-level intentional layers embodying goals relatively detached from and invariant to environmental change, and

[3] Directly contra Miller et al.'s (2021) recommendation to relax "rigid" associations for long-term "well-being" optimizing for "happiness" in the near-term, note that the present paper works from a teleological understanding of happiness, Aristotelian purposeful rather than pleasant, reinforcing the point that trading principle for personal security may not be of significant value. Summarily, where Miller et al. propose that agents seek slopes for informative error reduction via externally sourced "affordances" in potential self-realization, their account reflects dynamics associated with posterior DMN dynamics but neglects the internal slope corresponding with the metaphysical self as global attractor with corresponding affordances "self-affordances" associated with anterior DMN (particularly dorsal medial) dynamics as developed in the present paper.

with development and ongoing refinement of subservient throughput operations associated with personality development. Whether rendered in terms of enactivist skillful attunement, shortcut throughput layers in human beings or predictive codes passing messages down through computational hierarchies in biologically inspired neurorobots, such PC inspired models minimize error of fit to environment as agents "attune" themselves through enacted prior embodied anticipation in the perceptive enhancement of control over internal (embodied) and external environment (together, G&C's "situation"). Exercised in the reduction of disease and death inducing stress evidenced in the dissolution of higher-order goals due allostatic overload per G&dK (2021a), self is revealed in the breakdown (in robots, see Tani, 2017, on self-organized criticality and minimal self). G&dK associate resistance to "pruning" of such higher-order processes with "maturity" of bowtie goal hierarchies, drawing into question when and why such pruning is appropriate. When should such processes be dissolved to ensure bodily integrity, or retained through crippling stress in service of progress towards the goal states that they represent, e.g. by attuning to a "new" normal or acting from moral principle, "autonomously" (a capacity in the exercise of which G&dK, 2021a, p. 281, suggest that robots may excel; cf. White, 2020; 2021)?

With change in response to shifting environmental demands associated with Wozniak's "me" and context invariant goals embodied at higher levels of compression of G&dK's bowtie, context-dependent action proceeds via throughput at relatively lower levels, raising the possibility that there might be a sense of self apparent as higher-order throughput potential is not exercised, e.g. "I could do more," or remains yet underdeveloped, e.g. "I can do more," or which most poignantly denies immediate throughput in light of such potential, non-reflectively as an aspect of the embodied situation that is not context dependent, e.g. in the form of conscientious objection, "I will not do that"? A positive answer to this question points to a possible sense of self accompanying each intermediate "me" as one of how context-dependent instances of objective self-determination contribute to or impede actualization of highest levels of a goal hierarchy, aspirations in the realization of which we may associate with so-called "metaphysical" self. This possibility is explored, now.

In metaphysical self, briefly, we are looking for an invariant self-relation across levels of organization from immediate non-reflective to universal moral principle. How might such self-relation manifest in a human being? Some information is available, that the systems in question self-organize in the reduction of uncertainty and with it computational costs associated with tracking unnecessary variables thereby incurring excessive metabolic costs and with this allostatic overload, disease and death. What is necessary is thus a constraint on computation in the service of active inference over the timescales essential to the target architecture, binding personal, intersubjec-

tive, social, cultural-historical and relatively invariant principled moral levels, with such constraint answering to the metaphysical "I" as constant and pervasive, both in a realist neurobiological and in a phenomenal sense of always and already accompanying any given instant of self awareness at more intermediate levels.

Consider in this context Ortega y Gasset's (2002) characterization of self as a constant and pervasive phenomena in terms of "vocation" involving the sense of a globally orienting purpose in life (p. 135). Consistent with the preceding PC inspired review, for Ortega y Gasset (OyG), life is future-oriented, a purposeful self-seeking "program" in pursuit of a target state, "one's life's global project" that also serves as the source of value as objects and others either assist or hinder this pursuit (OyG, 2002, note 149, p. 214). Differently from Metzinger's minimal envelope, experience of one's global project is both essential to and fundamentally directed for OyG, presenting as "pressure" on the "evergoing determination of my present […] exerted on it by my future, i.e. by my vocation or what I have to be, whether I succeed in carrying it out or not (even in part)." (note 158, pp. 215)

Where might such a global project self arise and corresponding phenomenology be grounded in human beings? The default mode network (DMN) stands out as a candidate as it integrates past (memory, hippocampus and related areas) and future (project situations, frontal cortex and related areas) in purposeful imagination of possible situations ("complex goal-directed … memory-based simulations") (Schacter et al., 2012) and in autobiographical memory (Spreng et al., 2009). The "default mode network" was originally so called due to observed suppressed activity during task engagement, with greater suppression during more difficult tasks, and with increased activity in non-action contexts, e.g. mind-wandering. Early research characterized DMN activity as an aspect of shifting action across different contexts, with such activity consistent with recent enactivist accounts of "real-life skilled behavior" in terms of "metastable attunement" as suggested in section 2 of this paper, for example. More recent research has investigated task-related activity in the context of self-appraisal from childhood to adulthood, finding less activation of the anterior DMN especially the dorsal medial prefrontal cortex (dmPFC) during explicit self-appraisal with increasing age corresponding with self-development over the human life course as an aspect of increased functional segregation of anterior (future project) and posterior (actional) components of the DMN, concluding that reduced connectivity correlates with developing self-concept (Davey et al., 2019). One idea here is that implications of instantial self-determinations require less prospection, as expectations are established through prior routine interactions exercised during the life course, as self-concept stabilizes with experience, consistent with the execution of OyG's program as described above and in a process that we may associate with G&dK's (2021a) "maturity."

Adolescent development of the DMN also involves increased segregation from task-positive network activity during a period when cortical potential is highest, decreasing with adult myelination (Park et al., 2021; cf. Vandewouw et al., 2021) i.e. with maturation. Phenomenology characteristic of this development includes accounting for one's self as a social project for an "imaginary audience" in the construction of a personal "fable" (Buis, Thompson, 1989). Narrative self development has been considered the "highest form of cognitive integration" (Hirsh et al., 2013) with "trouble" in the form of challenges to personal convictions a defining aspect thereof (Bruner, 1997). Interestingly, challenges to "protected" values correlate with DMN activity (Kaplan et al., 2017). Recent research distinguishes between two dissociable DMN subsystems, one associated with "valence" and value, and another with "vividness" and detail of prospective (imagined, possible) situations, concluding that the construction of situations (from memory) and their evaluation as worth seeking are neurocognitively separable processes (Lee, Parthasarathi, Kable, 2021; cf. Pezzulo et al., 2021; also, the inchworm and bivalve model of White, 2014). Finally, distinguishable "conservative" and "disruptive" processes modulate development of lasting brain-wide DMN connectivity during adolescence (Vasa et al., 2020). Together, a relatively radical reconfiguration of the whole brain system is experienced including the rapid growth of the prefrontal cortex (and associated mirror systems) responsible for projections over increasingly distant time scales (Fuster, 1989; cf. Pujol et al., 2021). It is worth noting that increased segregation of DMN and task-positive subsystems during adolescence is associated with higher intelligence (Sherman et al., 2014). Indeed, overemphasis on learning engagements with the immediate object environment in education may be undesirable for human childhood development, as this separation may be inhibited (Immordino-Yang et al., 2012).

The proposal here is that the differentiation of developing frontal areas during adolescence from processes embodying action routines and value associations adopted during childhood potentiates the development of a relatively detached, globally orienting project future self-situation. This proposal is complementary to contemporary work in embodied cognition on development of self and consciousness in the context of PC and related approaches. For instance, Anna Ciaunica and colleagues suggest that embodiment within another body during gestation constitutes an "original prior" constraining ongoing development of the organism. On this view, gestation serves to prepare the developing organism for "co-homeostasis" during dependent childhood. And, the present view adds to Ciaunica and colleagues' view that adolescence represents an equally necessary stage wherein individuality emerges in the projection of a uniquely embodied project self-situation (cf. Ciaunica et al., 2021; Ciaunica, Safron, Dellafield-Butt, 2021).

On the present view, life as a global project "I" emerges through more or less normal development of especially the valence associated subsystem of the DMN as a more or less clearly conceived sense of purpose to realize these values in routine interaction with the social and objective world, and around which contextually specific, task-positive subsystems thereafter develop and are organized. Practically, each phenomenal "me" enacted during particular recurrent contexts in life such as when acting as a researcher, a family man, taking care of children, or exercising in a gym, can be represented as a set of sub-attractors of the DMN (corresponding with various bundle accounts surveyed in section 3, above). Valence (answering why these operations are worth performing and refining with increasing precision through directed epistemic agency over the life-course) binding these together develops as a global project "I" that can be characterized as a global attractor with the corresponding sense of self as purpose in life emergent as target valuations segregate from perceptual reality during segregation of developing highest-level default mode from task-positive neural processes.[4] Summarily thus, target state conditions embodied in these processes present as a life-long global project to bring the perceived reality in line with project values, with the felt tension between beginning and end situations accounting for phenomenality answering to Wozniak's "metaphysical 'I' " as well as to OyG's pressure on the present from the future. And, different senses of "me" emerge (including common uses of "I" that Wozniak would classify "me") as each uniquely situated self-seeking program is executed in circular interaction with the shared, objective world, towards embodied project ideals.

OyG's "vocation" answers to Wozniak's metaphysical "I" in the sense of a "lifelong persistent stream of consciousness" as it realizes aspects of itself as component instances of episodic "me" through interaction with a more or less undetermined and under-controlled world. This self-consciousness is not limited to the immediately embodied situation including conformity to social norms, and rather extends across representational time-spans to include invariant values and universal moral principles. Recalling G&C's (2020) relational self-positions within the scope of G&dK's (2021a) evolutionary goals, OyG's greater philosophy emphasizes that each individual occupies a privileged perspective on the shared world with unique potential to contribute to its ongoing determination as a common project through communication of personal experience, making history. The execution of this program, as such, is not a process that is reducible to activities arising between some fractions of brain activity such as might be the case with Wozniak's "delusion" or even necessarily within the confines of an individu-

[4] This is the slope for informative error reduction missing in Miller et al.'s (2021) account of "happiness" as global attractor.

al organism. Rather, OyG's vocational self seems to represent that constant aspect of self brought forward in G&dK's concept of higher-level goals established by an evolved "active inference engine" (G&dK, 2021a, p. 260) amongst other evolved active inference engines with similarly embodied aims.

Interesting in this context, Jesse Bettinger and Timothy Eastman (2017) consider biological cognition in terms of anticipation of self characterized as "predictive model space" that is "counterfactual" in the sense that "the model is an imperfect model trying to optimize its predictions and learn about the system it is modeling" (p. 114). The idea is already a familiar one, that cognition is essentially anticipatory, depending on established neural processes to respond to perceived reality—"information is encoded through synaptic weighting, and the confidence (or precision) of predictions can be altered by hierarchical gain modulation operating as generative models of the system regarding incoming sensory data" (Bettinger, Eastman, 2017, p. 112) — and preparing for most likely outcomes, with "predictions" being "contingent on actual sensory data to become active." (p. 114) Reviewing Alfred North Whitehead and the notion of "proposition," Bettinger and Eastman (2017) distinguish between "prehending" (perceiving) subjects and "logical subjects" in a way reflecting Wozniak's distinction between "me" and "I" mapped onto the perception-action cycle, with "me" upstream and "I" down consistent with preceding discussion (especially of L&F in section 3). On this account, prediction error is fed upstream, becoming the phenomenal "me" while the "I" is characterized as a "might be" on the propositional model of putting forward possibilities (predictions) towards which the living system then pulls itself through action "to maintain a inner-range of state values" evidencing "future-to-present (syntopic, attractor) logic" and apparent "backwards-in-time causality" in contrast with non-living physical system dynamics characterized in terms of "usual past-to-present" efficient causation (Bettinger, Eastman, 2017, p. 117–118), reminiscent of OyG's vocation.

In a way, Bettinger and Eastman capture the intention of the present proposal, with metaphysical self held out as a position to be realized through lifelong self-development. When asking "to what end" such anticipatory systems form, they answer to fulfill "existential needs before those needs become a crisis" (Bettinger, Eastman, 2017, p. 115, quoting Coffman, Mikulecky, 2015) and explore the roles of the salience network in conjunction with midline structures to constrain attention in the exercise of control via allostasis, but limit discussion to retention of individually embodied biophysical integrity. The present view considers the life-course of the organism as modulated by adolescent development as essentially propositional in that a global project ideal is embodied that thereafter serves to constrain cognition to temporally extended values in solution of evolutionary prob-

lems confronting the organism population as a whole, with such ideals informed by development of mirror neural systems thus shaping project ideals in a way that these may embody values independent of individual bodily integrity e.g. principles worth dying for, altruistic aims, rather than local, context-dependent attractors.

Finally, with this comparative account, Wozniak's challenge to "prove that there is a qualitative difference" and "to demarcate the exact border" (Wozniak, 2018, p. 12) between metaphysical subject and its ongoing iterative self-determination can be answered. The image of metaphysical self as essentially propositional places a forward project ideal against more immediate lower-level throughput processes satisfying the stipulation for an invariant self-relation with which this section began. It associates Wozniak's "I" with the feeling of being always and already in the context of progress towards defining values, as a program working to solve what is essentially itself as a uniquely embodied potential solution to evolutionary problems by bringing the perceived reality in-line with project ideals. Ongoing cognition on this model involves testing counterfactuals "as if" actually embodied self-positions (cf. Bettinger, Eastman, 2017; G&C, 2020) in resolution of this embodied project, effectively bridging inherited situations with ideal end states as moderated by adolescent development. Here again, it is important to emphasize the functional segregation of valence and vividness subsystems. With progress towards embodied project ideal, self is objectively determined, and contextually specific "me" related accessible details are embodied with associated processes maturing through iterative interaction toward this end (in this way answering to Wozniak's intuition that PC inspired inquiries into self might focus on so-called "access consciousness" as accessible details are encoded in the context of this metaphysical self-pursuit; cf. Davey et al., 2019).

## 5. DISCUSSION

With OyG's global project, we have the constant and pervasive sense of self answering to Wozniak's "I" that is objectively determined as an embodied inference engine engaging in "what if" processing over self-delimited predictive model space. Why should such higher-level goals develop, outstripping given contextual demands? The idea is that the metaphysical self as project ideal self-situation develops in response to emerging threats to evolutionary goals at levels of organization beyond the individually embodied agent and extending to all similarly embodied (here we may follow Kant in saying "rational") agency not necessarily limited to human agency but deriving from a similar process of adolescent development in other living systems, also (cf. Ledoux, 2021). Accordingly on the present view, self is

essentially purposeful, extending over the course of an anticipated life-span with the potential to represent target situations which an agent may not realistically anticipate inhabiting, e.g. Kant's Kingdom of Ends, though orienting intermediate action towards such ends, regardless.

The notion that the metaphysical self emerges as a globally orienting project binding current with ongoing and future actions in iterative self-determination towards a final project self-position, embodied as a unique proposition that "might be" a solution of evolutionary problems, and that forms during adolescent (highest-level) neural development, allows us to revisit the image of the bowtie. Emphasizing the temporal binding between perceptual instances, we may consider a "traveling bowtie" as one that binds perception and action in a nested goal hierarchy with throughput traveling across levels of processing according to contextual demands and with self actualized in this process. Moreover, consider in this context the heterochronic development of human beings i.e. from prefrontal to hippocampal processes as mediated by the thalamus perhaps specifically the reuniens nucleus (cf. Smaers et al., 2017), alongside changing default and task positive network connectivity, adding another dimension to the journey of the traveling bowtie as it matures over the life-course.

Recalling G&C's (2020) analysis in this light, a "realist" view of something like Wozniak's metaphysical self is constant in a way that is not captured in the traveling bowtie and the context-dependent bundle-theories that it represents. Rather, the "I" per OyG's vocation shapes experience regardless of context. One candidate grounds for constancy in the face of contextual change exists in L&F's (2018) "confidence" in project predictions. Confidence, though phenomenally context-dependent, is constant in that it always involves holding current alongside other, potentially embodied situations, in a way consistent with G&C's self-positional and Bettinger and Eastman's propositional selves, with agents constantly coming to terms with changing situations in a bottom-up and top-down manner (cf. White, 2010, 2014). Another empirical approach which touches on the omnipresence of metaphysical self is available in Andrew and Alexander Fingelkurts and Tarja Kallio-Tamminen's (Fingelkurts, Fingelkurts, Kallio-Tamminen, 2020, 2021) characterization of "witness consciousness" which, like the present view, draws on interplay of subsystems of the DMN. What is absent from witness consciousness is the sense of orientation and with it purpose and source of meaning in life, as with Metzinger's minimal envelope. Absent from L&F (as well as Miller et al., 2021) is how drive to minimize uncertainty informs OyG's vocational call to order apparent disorder from one's unique place in history thereby becoming a solution to G&dK's evolutionary problems through more or less freely directed development of personal potential.

How might such orientation to work, increasing order at personal expense, be best compressed and communicated? Consider the model of

a Platonic cave in complement to the bowtie model architecture, as it articulates a similar input-throughput-output dynamic while capturing the constant orientation to act towards the highest potential of OyG's vocation, making explicit the uncanny sense of self more or less present during routine short-circuits characteristic of metaphysical self according to the present view. There are three sections to the model. The cave represents routine conformity to social norms. The mountain above the cave represents one's highest potential self-situation in the representational space of ideals i.e. Mount Olympus, home of the Gods, corresponding to embodied global project per OyG. And, the reflecting pool on the plane between them under the shade of a tree affords a view of one's self as an object of reflection in front of the mountain behind that "me" representing one's highest calling, which together represent cognizance of purpose in life to achieve that project ideal and satisfy the judgement of the Gods. Ascending and descending the cave corresponds with input-throughput-output on the bowtie model, with the "I" perceived in the difference between the current reflected "me" and who one must become through one's life's work per OyG.

When gazing into the reflecting pool, "I" see "me" recalling Wozniak's resurrection of Wittgenstein's mirror. At the same time, changing focus I can see the mountaintop looming above the surveilled object "me" and against the ideals of which I can feel the space of my progress in self-development towards this highest aspiration. Recalling OyG's pressure from the future on the present, to become what I need to be or fail, one can imagine that it pulls the eyes upward and away from the downward gaze in the direction of the cave which orients agency in norm-seeking and exercising embodied routine per G&dK's short-circuits, effectively toward procedural self-unconsciousness through sufficient precision. G&dK's "personality development" may be seen to involve the balance of forces in either direction, with "maturity" involving the pruning of project ideals cognizant of the homeostatic cave environment in the minimization of stress. On Plato's account, the philosopher who represents this pressure to look upward, communicates the potential above the cave basin and reminds the slaves of their inner duties to seek their true vocations through action, is poorly treated by norm defenders, as the reminder of neglected higher-order goal-states causes stress, as if the cave environment were the one worth seeking, after all (cf. Miller et al., 2021). Here, we may compare Martin Heidegger's "fallenness" to the cave-bound condition, with sense of metaphysical self revealed in the call of conscience that he associates with philosophy (Heidegger, 1998).

The sense of self as outstanding, as propositional, and as represented by the distance between reflected "me" before the mountain-top of one's project ideal, is not captured on the bowtie models. The movement up and down from cave to reflecting pool may be captured by the traveling bowtie. But, the metaphysical self is the view on the present from the summit of

personal potential, OyG's pressure from the future, which on the bowtie model may be represented in the information passed upwards through a goal-hierarchy beyond short-circuit throughput layer, and downwards in comparison of current action against project ideals, delivering OyG's sense of self in the feeling of who "I" must become through a life of directed self-development. This sense of self is captured in the myth of the cave, in the image of the mountain rising above the introspected access of the reflecting pool. Moreover, this image indicates a deficiency in G&dK's assessment of self as relative stability with resilience in terms of maturity, as it allows a focus on the relationship between currently realized or anticipated and highest level project ideals perceived as the pull of OyG's vocation most obvious when one's global project is contravened. The cave model thus affords a focus on metaphysical self as the difference from norms, not realized in norm-seeking entrainment, but in norm-breaking autonomy, instead.

Recalling the situational self of G&C and other multistable routine bundle theories, cave-life represents routine enaction unaware of guiding norms with active and affective mirroring keeping cave inhabitants commonly oriented toward shadows projected by slavers on the cave wall. Given such a shared situation orchestrated for the benefit of others (e.g. politicians, global economic cabals), we can imagine minds "resonating" in coordinated action without prior planning, self-organizing in the common representational space (cf. Pöppel et al., 2021). Yet, there is a reason why this description of life is intuitively unattractive; it is self-nullifying. Far from evidencing a sense of self, the cave model illustrates the loss of self as a standing-out from routine and established norm. It is conceivable then that perceived instability given certain (social) situations is not evidence of mental or personality disorder, at all, and rather that it points to the existence of an outstanding sense of self from which a subject feels a frustrating alienation in the face of especially social pressures to conform to norms that contradict invariant values. This is to say that self as a proposition, in its present situation, is impeded from progressing toward its project conclusion, and the subject may experience debilitating anxiety in the dedication of metabolic potential, trying to compute a way out of the cave if not for one's self then for everyone who stands to suffer for the sub-optimal situation.

In Plato's allegory, a slave may twist at her or his bindings to catch a glimpse of something outside of the play of shadows, to see something of the slavers and their useful idiots who project their propaganda on the cave wall from above. Bound without hope of freedom to seek one's project self-situation, desperation and disorder may result. On this picture, self as project presents in the felt difference between established norms and project ideal with the tension of the chains experienced as stress. This sense of self is directional in that it pulls away from routine expectation, motivating the slave to break from habit, and reclaim one's self from the "they" of

a Heidegger or the "herd" of a Friedrich Nietzsche or the mass psychosis of a Mattias Desmet. This view is also complimentary to Kant's on personality, evident when action towards highest values runs contrary to established routine, and on whose account duty to one's self is experienced as a felt pull upward toward moral perfection, characteristic also of OyG's vocation.

It is in this potential to break free from habit, to stand out from norms of expectation, and to moreover communicate this potential to others who are somehow bound to less, that we may most directly associate the metaphysical "I." So, rather than in seeking resonance with established norms and contemporary expectations, we may identify the feeling of being an "I" in discord, for instance in conscientious objection and the power to say "no" through civil disobedience, extending to construction of popularly accessible accounts in the forms of myths and moral exemplars who die rather than act contrary to highest-order guiding principles, e.g. Martin Luther King Jr., Socrates, Christ. Here, we may offer a word on Jeffrey White and Jun Tani's (2016, 2017) notion of "myth consciousness" originally introduced in the context of cognitive neurorobotics. In that work, "most consciousness" can be associated with Wozniak's phenomenal "me" whereas "myth consciousness" represents awareness of being a metaphysical "I" in ongoing self-development toward ideal situations at most invariant levels of organization, "embodying history in all of its determinations." The cave model represents such a condition.

## 6. RELIGION, AFFORDANCE AND VALUE ALIGNMENT

The traveling bowtie model emphasizes the temporal binding associated with anticipation and prediction, but falls short of shedding light on integrative life-long self-development towards highest-level goal states extending past present personal and social constraints as informed by invariant moral values. This dynamic is captured on Plato's cave model, including also the sense that one has a duty to moral perfection, as in Kant, and the corollary that social norms represent the avoidance of this duty, as in Heidegger's inauthentically "fallen" condition i.e. hiding from one's highest potential behind idle chatter and other distractions characteristic of life in the cave.[5] Self-reflection affords a view on this highest potential, demanding freedom to pursue it through directed self-development—escaping from enslavement to shadows in the cave—as articulated by Plato with his reflecting pool.

[5] Fallenness is natural as routine enaction is not necessarily performed in avoidance of highest duties to self; usually, it is necessary, and a condition which may be associated with G&C's situational self alongside ecological enactivist accounts and other "bundle theories" as surveyed in the present paper. Inauthentic fallenness involves active neglect of highest potentials, including for example composition of academic, e.g. enactivist papers excusing foregone purpose through lack of resolve, the potential for which is not captured on any of the surveyed views.

Clarity on such dynamics affords brief consideration of the practice of prayer and the purpose of religion. Prayer can be viewed as a directed, meditative reinforcement of highest-level goals and iterative increase in precision of ongoing determination of global project-ideal self-situations including inventory of current self-position (in the sense of G&C) relative to project self-position (in the sense of OyG's vocation). This characterization naturalizes prayer as an affirmation of prospects put forward by evolved inference engines as embodied propositions. Prayer on this view can be appreciated as confirming the sense of metaphysical "I" deflated away on purely analytic approaches which fail to capture the intuitive sense of obligation to morally optimal outcomes common to human adolescent self-reports. Again, project self on the present view is a self-organizing solution to evolutionary problems, during and after development felt in the variable commitment to directing personal potential to overcome obstacles to evolutionary goals potentially including highest-possible human situations writ large, i.e. those associated with invariant values and universal moral principles e.g. Kant's Kingdom of Ends as Heaven on Earth. Thus, prayer as a practice of entrainment to evolutionary goals can be seen as prosocial and beneficial to the population of agents across generations not limited to the self-sacrificing Saint or other moral exemplar including potential artificial agents engineered with such capacities in mind (cf. Goekoop, deKleijne, 2021a, p. 281). Ultimately, there is nothing that seems to stand in the way of formalizing such processes, with robot religion providing a computational model proof-of-concept for the importance of faith in human beings (cf. White, 2021).

Here, some note is appropriate regarding moral consideration of artificial agents engineered according to the model of religion as entrainment to prosocial purposes sketched above. Vincent Muller and Michael Cannon (2021) distinguish between context specific "instrumental" and "general" intelligence, and consider that a "superintelligent" general AI may pursue any goal, possibly deviating from human goals, generating the "value alignment problem." Their account proceeds from a decision theoretic characterization of intelligence as a matter of maximizing expected utility, following Stuart Russell (2019). On this view, machines are engineered to optimize performance in specific operational contexts according to reward functions. Concerning current and anticipated technologies engineered accordingly, Muller and Cannon argue that potential value alignment problems derive from human rather than from AI initiatives. Though their distinction between general and instrumental intelligence is interesting as it can be roughly correlated with the functional orthogonality of value and vividness neurosystems considered in this paper, their treatment of "like us" AI reduces to variably broad operational contexts, neglecting value-orienting processes emphasized, herein (consider in this context results of Lee et al., 2021). Instead, their treatment reflects the enactive view from which they build and corre-

sponding characterization of DMN network functionality in terms of multi-stable norm-seeking (as introduced in section 3, above).

The position of the present work is that any intelligence "like us" undergoes different developmental periods (cf. Ciaunica and colleagues' gestation, adolescent development) and that in the process a uniquely embodied ideal goal-state self-organizes in highest-level neural processes that thereafter orients context-dependent engagements according to project values over the life-course. How this project is shaped determines to which values the agent thereafter strives, and what it takes as an opportunity for rewarding action. This view differs from for instance ecological enactivism, as to account for metaphysical self as proposed herein may require a radical revision of that position's Gibsonian take on affordances. Rather than nascent in the environment presenting to clever exploitation, on the present view affordance is better characterized as essentially "self-affordance" because the self as project exposes any genuine opportunity for progress towards its own ideal end as possibly mediated by external-environmental, ecological, factors (cf. Uexkull, 2010). Action motivated otherwise may run contrary to uniquely embodied goals, and so, though an opportunity for (perhaps expedient, norm-satisfying, stress-minimizing) action nascent in the environment, fail to be of meaningful value. To co-opt a popular example, so-called "higher-order" cognition employed to catch a bus to a job in which one is treated disposably by selfish men in the service of short-sighted vision, e.g. money through fraud, is not an opportunity for integrity-preserving action, and rather a chain of enslaving norm. It is not clear thus how ecological enactivists in particular can accommodate the present view without wholesale revision of their position, relocating focus to the internal environment and self-model away from e.g. architectural pre-occupations, relaxing principles for passing pleasure.

Considering "like us" AI in this context invites discussion about freewill in robots and recognition of rights typically afforded human beings on presumption of such potential. Vincent Muller (2021) argues that there is no need to consider robot rights, as contemporary model agents lack freewill and with it a sense of moral responsibility. With no individual locus of responsibility to serve as "bearer of moral status" such agents cannot be afforded rights. Similarly, Keith Farnsworth (2017) argues that freewill requires self-determination understood in terms of organizational closure (being a "Kantian whole") with an internal means for choosing among (more or less available) options according to an agent's "master function," and Rochat (2019) offers a complementary Kantian view that a sense of self-unity as organized and distinct from others is fundamental to any possible learning and experience (cf. Ciaunica, Safron, Dellafield-Butt, 2021). Farnsworth argues that contemporary artificial agents are not "Kantian wholes" in this way, so do not have freewill and with it moral responsibility, thereby

supporting Muller's view on robot rights. For Farnsworth, the master function of biological models is reproduction, ostensibly inconsistent with the present view of metaphysical self which involves pursuing opportunities for action towards an internally self-projected ideal goal-state which may have little to do with biological reproduction. The present view is that Farnsworth's "master function" is directly comparable to OyG's vocation, with the proposal here being that such developmental processes may be formalized for artificial embodiment in the foreseeable future, with robot rights considered accordingly. And, White (2021) argues directly that robots constructed on a Kantian model will be afforded comparable rights when this result is achieved.

## 7. CONCLUSION

The purpose of this paper has been to clarify metaphysical self in the context of contemporary predictive processing (PP) and predictive coding (PC) inspired accounts, to propose possible neural bases for its biological development, to expose how underlying structural dynamics may be represented, and to explore some of the implications of the view for ongoing work in different areas. In the survey of complementary accounts, Metzinger's minimal phenomenal experience (MPE) as the non-"egoic" "natural state" of an agent "predicting itself into existence" (note 26, p. 38, quoting Friston) was challenged. In regards to MPE, Metzinger (2020) concludes "the question of whether and in what sense it can count as 'fundamental,' and whether it is the only truly minimal state of consciousness, has not been answered" (p. 38). The view developed in response is that Metzinger's formalism represents neurosystem dynamics in especially human biological models, providing a kind of analytic envelope, but that the minimal conditions for self experienced as the target sense of a metaphysical subject "I" demands that his MPE envelope be extended in direction of an orienting self-project constraining and directing cognition (cf. Williford et al., 2018). We may consider here the image of a letter composing itself as it delivers itself to its propositional end, with the address on this letter emergent through developmental dynamics during adolescence in human beings. Metzinger's abstract envelope as revealed through meditative practice by developed adults with matured goal hierarchies is fundamental in the sense that it describes (through interoceptive access) the dimensionality of embodied processing associated with enaction, yet it fails to capture that process fundamental to the sense of self corresponding with Wozniak's "metaphysical 'I'" that can be associated with the address on an envelope in transit. In the case of AI, the view here is that such a sense of self and purpose as source of meaning may be formalized in recurrent neural networks constrained by different time-

scales at different levels, with higher order processes modeling increasing invariance associated with context independence and constancy such as in the case of moral principles, and with project aim emerging through developmental processes modeled after those embodied during adolescent development in human beings.

Finally, stepping into the context of current events, it is hoped that this work affords some clarity on the development of sense of purpose and meaning in life for contemporary young people. Recent OECD polling suggests that many adolescents report deficient sense of purpose (OECD, 2019) with nearly half of polled UK adolescents reporting an unsatisfactory sense of meaning in life, for example, raising the issue of the role of education in development of such a sense. At the same time, sense of purpose in life is understood to be protective against developmental disorders and risk taking behaviors including alcohol abuse and sexual promiscuity (Gongora, 2014; Brassai et al., 2011). Given current events (e.g. mandatory masking effectively obscuring crucially mirrored expressions of affect), current interest in understanding human enculturation in order to rectify social injustice for instance in resistance to corrupted leadership (cf. Haslanger, 2019), as well as potential association with personality disorders such as relative instability given changing situations dependent on strength of association with most meaningful, higher-order neural processes that rapidly develop at this stage of biological maturation in life (recalling discussion of G&dK, 2021a, above), policy-makers must be made cognizant of purpose-affirming ends consistent with evolutionary goals. The youth of today are the leaders of tomorrow, tasked with the construction of order in the face of looming disorder at all and increasing levels of organization, from the self on up (cf. G&dK, 2021b). It is ill-advised for policy to run contrary to development of evolved highest potentials, as such would amount to an evolutionary short-circuit and loss of meaning as highest values are contravened, what Michelle Maiese considers "moral atrophy" (Maiese, 2021).

Looking back through history, predictive self-modeling across levels of increasing invariance over increasing time-scales as an aspect of embodied development has reached high-points in the visions of moral exemplars as represented in cultural and mythical heroes and Gods, for example. Pursuit of these ideals has delivered human beings to the present stage. Associations with these high-points and their expressed ideals are lasting, representing invariant values to which persons aspire as propositions and towards which they "predict themselves into existence," thus answering to White and Tani's (2016, 2017) myth consciousness in the felt potential to embody the space of history and all of its determinations. Prayer and some forms of meditation would seem to reinforce these connections, grounding justified resistance to oppression in the call to something greater. The value of such practices to these ends deserves more attention in future work, in the context of artificial

religion and value alignment in artificial general intelligence, and moreover may help to inform current interest of social scientists in ideological oppression with conformity associated with atrophied moral cognition and the correlate dimming of human potential.

## REFERENCES


P. Bagus, Peña-Ramos, J. A., & Sánchez-Bayón, A. *COVID-19 and the Political Economy of Mass Hysteria*, International Journal of Environmental Research and Public Health, 18 (4), 2021; https://doi.org/10.3390/ijerph18041376

J. Bettinger, Eastman, T., *Foundations of Anticipatory Logic in Biology and Physics*, Progress in Biophysics and Molecular Biology, 131, 2017, pp. 108–120.

O. Blanke, Metzinger, T., *Full-body Illusions and Minimal Phenomenal Selfhood*, Trends Cogn. Sci. 13, 2009, pp. 7–13; doi: 10.1016/j.tics.2008.10.003

N. Block, *On a Confusion about a Function of Consciousness*. Behav. Brain Sci. 18, 1995, pp. 227–247. doi: 10.1017/S0140525X00038188

L. Brassai, Piko, B. F., Steger, M. F., *Meaning in Life: Is It a Protective Factor for Adolescents' Psychological Health?*. International Journal of Behavioral Medicine, 18(1), 2011, pp. 44-51.

J. Bruineberg, Seifert, L., Rietveld, E., Kiverstein, J., *Metastable Attunement and Real-life Skilled Behavior*, Synthese, 2021; https://doi.org/10.1007/s11229-021-03355-6

J. Bruner, *A Narrative Model of Self-Construction*, Annal NY Acad Sci, 818, 1997, pp. 145–161.

J. M. Buis, Thompson, D. N. *Imaginary Audience and Personal Fable: a Brief Review*, Adolescence, 24 (96), 1989, pp. 773–781.

A. Ciaunica, Constant, A., Preissl, H., Fotopoulou, K., The First Prior: *From Co-embodiment to Co-homeostasis in Early Life*, Consciousness and Cognition, 91, 2021; https://doi.org/10.1016/j.concog.2021.103117

A. Ciaunica, Safron, A., Delafield-Butt, J., *Back to Square One: the Bodily Roots of Conscious Experiences in early life*. Neuroscience of Consciousness, 2021, 2; https://doi.org/10.1093/nc/niab037

R. W. Clowes, Gärtner, K., *The Pre-reflective Situational Self*. Topoi, 39(3), 2020, pp. 623–637.

J. A. Coffman, Mikulecky, D. C., *Global Insanity Redux*. Cosmos and History, 11(1), 2015, pp. 1–14.

C. G. Davey, Fornito, A.; Pujol, J.; Breakspear, M.; Schmaal, L.; Harrison, B. J., *Neurodevelopmental Correlates of the Emerging Adult Self,* Dev Cogn Neurosci, 36, 2019; https://doi.org/10.1016/j.dcn.2019.100626

G. Deco, Jirsa, V. K., *Ongoing Cortical Activity at Rest: Criticality, Multistability, and Ghost Attractors,* Journal of Neuroscience, 32 (10), 2012, pp. 3366–3375.

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

K. Farnsworth, C*an a Robot Have Free Will*?. Entropy, 19 (5), 237, 2017; https://doi.org/10.3390/e19050237

A. A. Fingelkurts, Fingelkurts, A. A., Kallio-Tamminen, T., *Selfhood Triumvirate: From Phenomenology to Brain Activity and Back Again*, Consciousness and Cognition, 86, 103031, 2020; https://doi.org/10.1016/j.concog.2020.103031

A. A. Fingelkurts; Fingelkurts, A.A.; Kallio-Tamminen, T. Self, *Me and I in the Repertoire of Spontaneously Occurring Altered States of Selfhood: Eight Neurophenomenological Case Study Reports*, Cognitive Neurodynamics, 2021, pp. 1–28; https://doi.org/10.1007/s11571-021-09719-5

J. M. Fuster, *The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe,* Raven Press, New York 1989.

K. Gärtner, Clowes, R. W., *Predictive Processing and Metaphysical Views of the Self*, The Science and Philosophy of Predictive Processing, D. Mendonca, M. Curado, S. Gouveia

(eds.), Bloomsbury: London, UK, 2020, pp. 59–81; https://doi.org/10.5040/9781350099784.ch-004

R. Goekoop; deKleijn, R., *How Higher Goals Are Constructed and Collapse under Stress: a Hierarchical Bayesian Control Systems Perspective*, Neuroscience & Biobehavioral Reviews, 123, 2021a, pp. 257–285.

___ , *Permutation Entropy as a Universal Disorder Criterion: How Disorders at Different Scale Levels Are Manifestations of the Same Underlying Principle*, Entropy 23, 1701, 2021b; https://doi.org/10.3390/ e23121701

V. C. Góngora, *Satisfaction with Life, Well-being, and Meaning in Life as Protective Factors of Eating Disorder Symptoms and Body Dissatisfaction in Adolescents*, Eat Disord., 22 (5), 2014, pp. 435–449; DOI: 10.1080/10640266.2014.931765.

S. Haslanger, *Cognition as a Social Skill*, Australasian Philosophical Review, 3 (1), 2019, pp. 5–25; DOI: 10.1080/24740500.2019.1705229.

M. Heidegger, *Being and Time*, State University of New York Press, Albany, N.Y 2010.

J. B. Hirsh, Mar, R. A., Peterson, J. B., *Personal Narratives as the Highest Level of Cognitive Integration*, Behav. Brain Sci., 36 (3), 2013, pp. 216–217.

J. Hohwy, Michael, J., *Why Should Any Body Have a Self?*, in: The Subject's Matter: Self-Consciousness and the Body, de Vignemont and Alsmith (eds.), MIT Press, 2017, pp. 363–391.

IM. H. mmordino-Yang, Christodoulou, J. A., Singh, V., *Rest Is Not Idleness: Implications of the Brain's Default Mode for Human Development and Education,* Persp. on Psych. Sci., 7 (4), 2012, pp. 352–364.

S. Kahl, Wiese, S., Russwinke, N., Kopp, S., *Towards Autonomous Artificial Agents with an Active Self: Modeling Sense of Control in Situated Action*, Cognitive Systems Research, 72, 2022, pp. 50–62; https://doi.org/10.1016/j.cogsys.2021.11.005

J. T. Kaplan, Gimbel, S. I., Dehghani, M., Immordino-Yang, M. H., Wong, J. D., Tipper, C. M., Damasio, H., Damasio, A., Sagae, K., Gordon, A. S., *Processing Narratives Concerning Protected Values: a Cross-Cultural Investigation of Neural Correlates*, Cerebral Cortex, 27 (2), 2017, pp. 1428–1438.

J. E. LeDoux, *As Soon as There Was Life, There Was Danger: the Deep History of Survival behaviours and the Shallower History of Consciousness*, Phil. Trans. R. Soc. B, 377, 2021; https://doi.org/10.1098/rstb.2021.0292

S. Lee, Yu, L., Q., Lerman, C., Kable, J. W., *Subjective Value, Not a Gridlike Code, Describes Neural Activity in Ventromedial Prefrontal Cortex during Value-based Decision-making*, Neuroimage, 237, 118159, 2021; https://doi.org/10.1016/j.neuroimage.2021.118159

S. Lee, Parthasarathi, T., Kable, J. W., *The Ventral and Dorsal Default Mode Networks Are Dissociably Modulated by the Vividness and Valence of Imagined Events*, J. Neurosci., 41 (24), 2021, pp. 5243–5250.

C. Letheby, Gerrans, P., *Self Unbound: Ego Dissolution in Psychedelic Experience*, Neuroscience of Consciousness, (1), 2017, pp. 1–11.

J. Limanowski, & Friston, K. *'Seeing the Dark': Grounding Phenomenal Transparency and Opacity in Precision Estimation for Active Inference*. Frontiers in Psychology, 9, 643, 2018, pp. 1-9; https://doi.org/10.3389/fpsyg.2018.00643

J. Limanowski, & Friston, K. *Attenuating oneself: An active inference perspective on "selfless" experiences*. Philosophy and the Mind Sciences, 1, 2020, pp. 1–16.

M. Maiese, *Mindshaping, Enactivism, and Ideological Oppression*, Topoi, 2021; https://doi.org/10.1007/s11245-021-09770-1

T. Metzinger, *Phenomenal Transparency and Cognitive Self-reference*, Phenomenology and the Cognitive Sciences 2, 2003, pp. 353–393; https://doi.org/10.1023/B:PHEN.0000007366.42918.eb

___ , *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. Basic Books: New York 2009.

___ , *Why Are Dreams Interesting for Philosophers? The Example of Minimal Phenomenal Selfhood, Plus an Agenda for Future Research,* Frontiers in Psychology, 4, 2013; https://doi.org/10.3389/fpsyg.2013.00746

___ , *Minimal Phenomenal Experience: Meditation, Tonic Alertness, and the Phenomenology of "Pure" Consciousness*, Phil. And the Mind Sci., 1(I), 2020, pp. 1–44; https://doi.org/10.33735/phimisci.2020.I.46.

M. Miller, Kiverstein, J, Rietveld, E., *The Predictive Dynamics of Happiness and Well-Being*, Emotion Review, 14(1), 2022, pp. 15–30; doi:10.1177/17540739211063851

V. C. Muller, *Is It Time for Robot Rights? Moral Status in Artificial Entities*, Ethics and Information Technology, 23, 2021, pp. 579–587; https://doi.org/10.1007/s10676-021-09596-w

A. Newen, *The Embodied Self, the Pattern Theory of Self, and the Predictive Mind*, Frontiers in Psychology, 9, 2270, 2018, pp. 1–14; https://doi.org/10.3389/fpsyg.2018.02270

N. Nyberg, Duvelle, E., Caswell, B., Spiers, H. *Spatial Goal Coding in the Hippocampal Formation*, Neuron, 2022; https://doi.org/10.1016/j.neuron.2021.12.012

OECD, PISA, *2018 Results* (Vol. III)*: What School Life Means for Students' Lives*, PISA, OECD Publishing, Paris, 2019; https://doi.org/10.1787/acd78851-en

J. Ortega y Gasset, *What Is Knowledge?,* State University of New York Press, Albany 2002.

B.-Y. Park, Paquola, C., Bethlehem, R. A. I., Benkarim, O., *Neuroscience in Psychiatry Network (NSPN) Consortium*, Mišic, B., Smallwood, J., Bullmore, E. T., Bernhardt, B. C., *Adolescent Development of Multiscale Cortical Wiring and Functional Connectivity in the Human Connectome*, BioRxiv 2021.08.16.456455, 2021; https://www.biorxiv.org/ content/10.1101/2021.08.16.456455v2

G. Pezzulo, Parr T, Friston K. *The Evolution of Brain Architectures for Predictive Coding and Active Inference*, Phil. Trans. R. Soc. B, 377, 20200531, 2021; https://doi.org/ 10.1098/rstb.2020.0531

J. Pöppel, Kahl, S., Kopp, S. *Resonating Minds—Emergent Collaboration Through Hierarchical Active Inference*, Cognitive Computation, 2021, pp. 1–22; https://doi.org/10.1007/ s12559-021-09960-4

J. Pujol, Blanco-Hinojo, L., Macia, D., Martnez-Vilavella, G., Deus, J., Prez-Sola, V., Cardoner, N., Soriano-Mas, C., Sunyer, J., *Differences between the Child and Adult Brain in the Local Functional Structure of the Cerebral Cortex*, Neuroimage, 237, 2021, 118150; https://doi.org/10.1016/j.neuroimage.2021.118150

P. P. Rochat, *Self-Unity as Ground Zero of Learning and Development*, Frontiers in Psychology, 10, 2019; https://doi.org/10.3389/fpsyg.2019.00414

S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking, 2019.

A. Safron, *The Radically Embodied Conscious Cybernetic Bayesian Brain: from Free Energy to Free Will and Back Again,* Entropy, 23 (6), 2021; https://doi.org/10.3390/e23060783

D. L. Schacter, Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., Szpunar, K. K., *The Future of Memory: Remembering, Imagining, and the Brain*, Neuron, 76 (4), 2012, pp. 677–694.

E. Sennesh, Theriault, J., Brooks, D., van de Meent, J., Feldman Barrett, L., Quigley, K., *Interoception as modeling, allostasis as control,* Biological Psychology, 167, 2022. https://doi.org/10.1016/j.biopsycho.2021.108242.

L. E. Sherman, Rudie, J. D., Pfeifer, J. H., Masten, C. L., McNealy, K., Dapretto, M., *Development of the Default Mode and Central Executive Networks Across Early Adolescence: a Longitudinal Study*. Develop Cogn Neurosci 10, 2014, pp. 148-159.

J. B. Smaers, Gomez-Robles, A., Parks, A. N., & Sherwood, C. C. *Exceptional Evolutionary Expansion of Prefrontal Cortex in Great Apes and Humans*, Current Biology, 27 (5), 2017, pp. 714–720.

R. N. Spreng, Mar, R. A., Kim, A. S. N., *The Common Neural Basis of Autobiographical Memory, Prospection, Navigation, Theory of Mind, and the Default Mode: a Quantitative Meta-Analysis*. J. Cogn. Neurosci., 21 (3), 2009, pp. 489–510.

P. Sterling, Allostasis, *A Model of Predictive Regulation,* Physiology & Behavior, 106 (1), 2012, pp. 5–15.

J. Tani, *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*, Oxford University Press: Oxford, UK 2017.

J. Tani, White, J., *From Biological to Synthetic Neurorobotics Approaches to Understanding the Structure Essential to Consciousness*, Part 2, APA Newsletter on Philosophy and Computers, 16 (2), 2017, pp. 29–41.

___ , *Cognitive Neurorobotics and Self in the Shared World, a Focused Review of Ongoing Research*. Adaptive Behavior, 2020; https://doi.org/10.1177/1059712320962158

J. Uexküll, *A Foray into the Worlds of Animals and Humans*, University of Minnesota Press: Minneapolis, MN 2010.

M. Valmisa, *What Is a Situation?*, in: Coming to Terms with Timelessness: Daoist Time in Comparative Perspective, L. Kohn (ed.), Three Pines Press, St. Petersburg, FL 2021, pp. 26–49.

M. M. Vandewouw, Hunt, B. A. E., Ziolkowski, J., Taylor, M. J., *The Developing Relations between Networks of Cortical Myelin and Neurophysiological Connectivity*. Neuroimage. 237, 118142, 2021; https://doi.org/10.1016/j.neuroimage.2021.118142

F. Váša, Romero-Garcia, R., Kitzbichler, M. G., Seidlitz, J., Whitaker, K. J., Vaghi, M. M., Kundu, P., Patel, A. X., Fonagy, P., Dolan, R. J., Jones, P. B., Goodyer, I.M., the NSPN Consortium, Vértes, P.E., Bullmore, E.T. *Conservative and Disruptive Modes of Adolescent Change in Human Brain Functional Connectivity*, Proc. Nat. Acad. Sci. U.S.A., 117 (6), 2020, pp. 3248–3253; DOI: 10.1073/pnas.1906144117.

J. White, *Understanding and Augmenting Human morality: An Introduction to the ACTWith Model of Conscience*, Studies in Computational Intelligence, 314, 2010, pp. 607–621.

___ , *Models of Moral Cognition*. In: Model-Based Reasoning in Science and Technology, L. Magnani (ed.), Springer, Berlin, 2014, pp. 363–391.

___ , *Autonomous Reboot: Aristotle, Autonomy and the Ends of Machine Ethics*, AI & Society, 2020; https://doi.org/10.1007/s00146-020-01039-2

___ , *Autonomous Reboot: Kant, the categorical imperative, and contemporary challenges for machine ethicists*. AI & Society, 2021; https://doi.org/10.1007/s00146-020-01142-4

J. White, Tani, J., *From Biological to Synthetic Neurorobotics Approaches to Understanding the Structure Essential to Consciousness, part 1*, APA Newsl. Philos. Comput. 16 (1), 2016, pp. 13–23.

___ , *From Biological to Synthetic Neurorobotics Approaches to Understanding the Structure Essential to Consciousness, part 3*, APA Newsl. Philos. Comput. 17 (1), 2017, pp. 11–22.

W. Wiese, *Explaining the Enduring Intuition of Substantiality. The Phenomenal Self as an Abstract 'Salience Object'*, Journal of Consciousness Studies, 26 (3–4), 2019, pp. 64–87.

K. Williford, Bennequin, D., Friston, K., Rudrauf, D., *The Projective Consciousness Model and Phenomenal Selfhood*, Frontiers in Psychology, 9, 2018; https://doi.org/10.3389/fpsyg.2018.02571

L. Wittgenstein, *Preliminary Studies for the "Philosophical Investigations", Generally Known as the Blue and Brown Books*. Blackwell, Oxford 1959.

M. Wozniak, *"I" and "me": the Self in the Context of Consciousness*. Front in Psych, 9, 2018, https://www.frontiersin.org/article/10.3389/fpsyg.2018.01656

ABOUT THE AUTHOR — PhD, Philosophy, University Missouri-Columbia, NOVA-LINCS (visiting researcher), Departamento de Informática, FCT/UNL, Quinta da Torre P-2829-516, Caparica, Portugal, and OIST (visiting researcher), cognitive neurorobotics research group, Okinawa, Japan

Email: jeffreywhitephd@gmail.com

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Eduardo Camargo, Ricardo Gudwin

# FROM SIGNALS TO KNOWLEDGE AND FROM KNOWLEDGE TO ACTION: PEIRCEAN SEMIOTICS AND THE GROUNDING OF COGNITION

doi: 10.37240/FiN.2022.10.zs.5

***ABSTRACT***

Cognition is meant as the process of acquiring knowledge from the world. This process is supposed to happen within agents, which build such knowledge with the purpose to use it to determine their actions on the world. Following Peircean ideas, we postulate that such knowledge is encoded by means of signs. According to Peirce, signs are anything that can be used to represent anything else. Also, for Peirce, to represent means to be able to generate another sign, called the interpretant of the original sign, which still holds the same power of interpretability, I.e, its power to be transformed into a new sign, holding this same power. This happens through a process called semiosis, the process by which a sign is transformed into an interpretant. This whole process is performed with the aim of subsidizing the agent in deciding its behavior. So, even though the semiosis process has the power to continue infinitely, it usually stops whenever the generated interpretant brings enough information in order for the agent to effectively act in the world. We take signals to be the substract of signs. Signals are any physical property, which can be measured and captured by the agent, by means of its sensors. This includes any kind of internal memory the agent is able to have access, in order to operate. In this sense, signs can be both in the world (if these signals come from sensors) and within the own agent's mind (if signals come from an internal memory). We understand an agent's mind as the agents' control system. In either case, signals can be abstracted as numbers. Not simply numbers, but numbers coming from specific sensors or specific memories. Using ideas from Peircean philosophy, in this work we postulate a pathway, in which signals, collected by either sensors or memory, can be organized in such a way that they can be effectively used as knowledge, in order for an agent to be able to decide its actions on the world, on the pursuit of its internal motivations. We postulate that agents identify and create a model of the world based on possibilities, existents, and laws, and based on this model, they are able to decide an action that maximizes the chance for the world to gain a shape, which the agents intend for it to be. This theory is postulated particularly for the case of artificial autonomous agents, meant to be constructed by engineering artifacts.

**Keywords**: Peircean semiotics, knowledge representation, cognitive science.

## 1. INTRODUCTION

Cognitive science is an interdisciplinary scientific field that connects diverse but related disciplines such as experimental psychology, theoretical linguistics, mathematical logic, and artificial intelligence with the aim of understanding how mind works. In this context, according to Bermudez (2020, pp.15-35), cognition is a form of information processing that allows organisms to interact with their environment to survive. This interaction involves a mind that is responsible for representing the world and developing knowledge about it, providing conditions for the organisms to explore and transform their environment. In short, cognition is meant as the process of acquiring knowledge from the world by an entity equipped with perceptive and actuation devices. Therefore, adopting the approach of *Embodied Situated Cognition,* proposed by Francisco Varela et al. (1991), we consider that perception and action instances are two strongly connected parts of the whole system (see Figure 1), and following some past insights from Gudwin (2014; 2015), we postulate in this work that the study of this kind of connection can be supported by the General Theory of Signs of the American philosopher Charles Sanders Peirce (*1839, +1914).

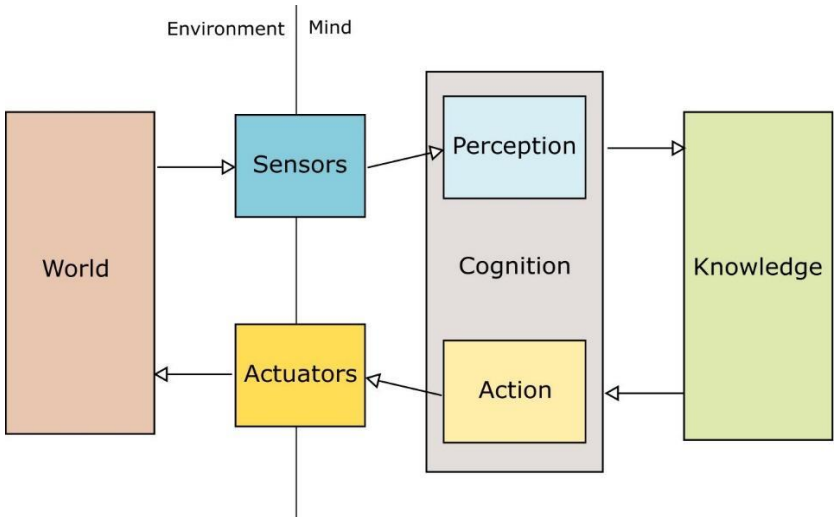



**Figure 1:** The scope of cognition based on the Embodied Situated Cognition movement

The concept of relations between an organism and its environment, mediated by a mind, takes in account that reality is organized as a system in which each organism affects the surrounding environment and is affected back by the things of the world, including other organisms. However, the presence of a mind must not be considered as an exclusive attribute of bio-

logical entities. As pointed out by Margaret Boden, "... the relation between life and mind is still highly problematic [...]. The common-sense view is that the one (life) is a precondition of the other (mind). But there's no generally accepted way of proving that to be so" (Boden, 2006, p. 1443). In accordance with a broader concept of mind, Peirce claims that thought and mind are not exclusively human attributes and must not be confounded with consciousness. For him, mind is a synonym of representation, and its actuation upon the matter occurs by force of certain laws of final causality. Thus, wherever there are laws, regularities and potentiality there is also rationality, and this should not presuppose consciousness but incorporated knowledge (Santaella, 1994). In Peirce's own words:

> "Thought is not necessarily connected with a brain. It appears in the work of bees, of crystals, and throughout the purely physical world; and one can no more deny that it is really there, than that the colors, the shapes, etc., of objects are really there. [...] Not only is thought in the organic world, but it develops there. But as there cannot be a General without Instances embodying it, so there cannot be thought without Signs. [...] Admitting that connected Signs must have a Quasi-mind, it may further be declared that there can be no isolated sign. Moreover, signs require at least two Quasi-minds; a Quasi-utterer and a Quasi-interpreter; and although these two are at one (i.e., are one mind) in the sign itself, they must nevertheless be distinct. In the Sign they are, so to say, welded. Accordingly, it is not merely a fact of human Psychology, but a necessity of Logic, that every logical evolution of thought should be dialogic." (CP 4.551)[1]

Peircean semiotics is a kind of phenomenology in which an interpreter's mind is affected by signals coming either from the world and/or from internal memories. The interpreter has no direct access to real objects (Dynamical Objects), but only to their signs conveyed by the signals (Immediate Objects), which means that all representation is due to some kind of collateral experience (CP 8.314). Peirce considers signs as anything suitable to represent anything else. Also, for Peirce, to represent means to be able to transform a sign into another sign called the interpretant of the original sign, which still holds the same power of interpretation. This happens through a process called semiosis, the process by which a sign is transformed into another sign. According to Noth, despite *sign, representation, mediation,* and *interpretation* being the key terms in the study of semiotic processes, instead of the term *information,* Peirce had much more to say about how signs convey information than is usually acknowledged in contemporary information sciences, and he explains:

---

[1] Citations to Peirce's works in this paper follow the traditional format used by Peircean scholars. So, instead of Peirce (1931–1958, pp. 120–138) for some pages of the *Collected Papers of Charles Sanders Peirce,* the citation appears as CP x.y where: CP indicates the title, x indicates the volume and y indicates the paragraph.

> "Peirce's information theory does not conceive of information in terms of probabilities of the occurrence of signals, words, or sentences in actual utterances. Instead of probabilities, it calculates the logical quantities of extension and intension of symbols. Furthermore, it does not only calculate the value of the actual information conveyed through new informative propositions but also information as it has accumulated through the implications that symbols acquire in the course of their history. It is, hence, both a theory of knowledge acquisition and a theory of the growth of symbols." (Noth, 2012)

So, if cognition means the process of acquiring knowledge from the world through a form of information processing, and signs convey information through semiosis toward knowledge acquisition, we postulate that knowledge is encoded through signs, which points to the Peircean semiotics as promising grounding for cognition. And finally, we consider here that the interpreting mind involved in the semiotic processes could be a computational device referred to as an artificial autonomous agent, or simply an agent which is defined by Russell and Norvig (2020, p. 36) as "anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators." Thus, in this paper, we consider the possibility of an agent, affected by the signals from the world, to use these signals in semiotic processes to acquire knowledge, and use this knowledge to drive its actions.

## 2. KNOWLEDGE AND CATEGORIES

The general theory of knowledge aims to understand the signification processes of human thought and their relations to the objects of the world as a whole. In its turn, the special theory of knowledge investigates the most elementary concepts used to describe objects, and these concepts are called categories. In this sense, the special theory of knowledge is a theory of categories, and its focus points to the logical origins of the forms of thought and how they arise from the essential laws of thought in confrontation with the experience data (Hessen, 2003, pp. 133–134).

Aristotle was the first philosopher to take care of such matters, and had used language, particularly classes of words, to frame all elements of reality into ten categories: substance; quantity; quality; relatives; somewhere; sometime; being in a position; having; acting; and being acted upon. According to Aristotle, words are things that are said, and it is natural to interpret his system as a classification of words. However he was not primarily interested in words but in the world to which words correspond (Studtmann, 2021).

Due to the skepticism about our capacity to distinguish precise divisions in reality, an important shift from Aristotelian realism to what was called

categorial conceptualism was promoted by Kant. Unlike his precursor, he established his system based on the idea that human thought has no access to *the thing in itself* but only to its appearance or *phenomena,* which leads to the essential categories that govern human understanding or judgement, and judgement is the basis for any possible cognition of phenomena. Thus, to enumerate the forms of possible judgement, he used Aristotelian logic to determine four respects in which one can classify any judgement: quantity, quality, relation, and modality. Moreover, in respect to each class of judgement, Kant recognized three subdivisions leading to twelve categories: Quantity (Unity, Plurality, and Totality), Quality (Reality, Negation, and Limitation), Relation (Inherence and Subsistence, Causality and Dependence, and Community), and Modality (Possibility, Existence, and Necessity) (Thomasson, 2019).

The third great system of categories in the history of philosophy was proposed by Charles Sanders Peirce. In his intent to describe the most universal and elementary categories of all possible experiences he followed the same terminology of Aristotle *(hai kategoriai)* and Kant *(die categorien),* but the result he achieved was even more radical than that of his predecessors. Based on the semiotic processes, he has found only three categories in which all phenomena can be divided, leading to a logical and social theory of sign (Santaella, 2000, p. 7). The following section presents a panoramic view of Peirce's work on semiotics.

## 3. BASIC NOTIONS OF PEIRCEAN'S SEMIOTICS

Semiotics denotes the study of *signs* and *significant processes* (semiosis). In modern semiotics, the general theory of signs of Peirce postulates semiotics as a universal science, not restricted to human communication. In such sense, signs do not correspond to a specific class of phenomenon but are the elementary components of a kind of phenomenology (Noth, 1995, pp. 39–41) or *phaneroscopy,* which aims to study the universal categories of the *phanerons* (from the greek *phaneros:* visible, manifest, evident, apparent):

> "What I term *phaneroscopy* is that study which, supported by the direct observation of phanerons and generalizing its observations, signalizes several very broad classes of phanerons; describes the features of each; shows that although they are so inextricably mixed together that no one can be isolated, yet it is manifest that their characters are quite disparate; then proves, beyond question, that a certain very short list comprises all of these broadest categories of phanerons there are; and finally proceeds to the laborious and difficult task of enumerating the principal subdivisions of those categories." (CP 1.286)

Peirce's efforts have reduced all phenomena to only three ontological categories:

— *Firstness* as "the mode of being of that which is such as it is, positively and without reference to anything else" (CP 8.328). Firstness is relate to the ideas of simple potentiality, possibility and independence, a feeling not yet converted to reflection, just a glimpse of reality in the state of pure indetermination. All ideas that are absolutely independent of further ideas to subsist are related to Firstness;

— *Secondness* as "the mode of being of that which is such as it is, with respect to a second but regardless of any third" (ibidem). Secondness points to the experience of space-time, to action, to the experiential reality, to fact, to a perceptible consistency without purpose or judgement, because all these ideas require a relation to other ideas in order to be conceived—any point in space or time requires a connection to another point in space or time in order to be space or time, any action requires an actor, a fact requires an existence where the fact materializes, a consistency requires a reference to what it is consistent to. This is the category for ideas that can only make sense while relating to another idea;

— *Thirdness* as "the mode of being of that which is such as it is, in bringing a second and third into relation to each other" (ibidem). Thirdness corresponds to mediation, to law or habit, to continuity, to purpose and judgement, to thought, and representation because all these ideas are, in themselves, relations between other ideas: mediation is the relation between two other things, a law puts under relation all its possible instances, a habit is nothing more than a learned law, continuity is the principle of recursive mediation between two others, purpose is a glimpse of the future mediating the flow of past to present. Thirdness is the category for ideas that are the own relation between two other ideas or, in other words, when the own relation of two other ideas becomes an idea.

For Peirce, the three categories are related in a triadic way and this relation is irreducible, without boundaries between them: "Not only does Thirdness suppose and involve the ideas of Secondness and Firstness, but never will it be possible to find any Secondness or Firstness in the phenomenon that is not accompanied by Thirdness" (CP 5.90). From this phenomenological framework, Peirce creates the notion of genuine sign, an ingenious explanation conceived from the concepts of thought and representation present in the category of Thirdness:

> "A Sign, or Representamen, is a First which stands in such a genuine triadic relation to a Second, called its Object, as to be capable of determining a Third, called its Interpretant, to assume the same triadic relation to its Object in which it stands itself to the same Object. The triadic relation is genuine, that

is its three members are bound together by it in a way that does not consist in any complexus of dyadic relations" (CP 2.274).

As pointed by Adele Queiroz (2004, p. 53), the tripod is the best representation of a sign. However, due to the purpose of making explicit some specific aspects of semiosis, this work uses a slightly modified tripod to represent the sign, but, despite its unbalanced form, no dyadic relation must be interpreted (See Figure 2).

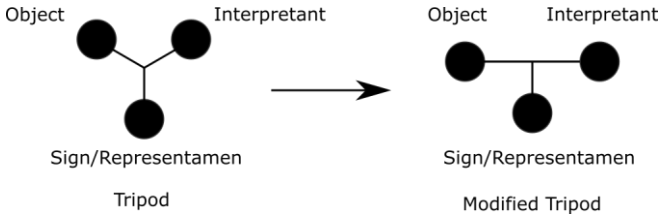



**Figure 2:** From the tripod to a modified tripod as sign representations

With our modified tripod, we want to emphasize that the sign/representamen, corresponds exactly to the relation between the object and the interpretant, detached as an entity in itself. Or, in other words, the sign is the third that connects the object to the interpretant, mediating between these other two, being the logic mediator between them.

Furthermore, Noth (1995, pp. 43–44) claims that Peirce considers two types of objects: the *Dynamical Object* corresponding to the object outside the sign, something near to the real object, and the *Immediate Object* corresponding to the object inside the sign, near to the representation itself; Moreover, Peirce also considers three types of interpretants, the *Immediate Interpretant* as a semantic potentiality, the *Dynamical Interpretant* as the direct effect produced by a sign in the interpreter which can lead to an action in the world, and the *Final Interpretant* as the one carried so far that an ultimate conclusion was reached, which signifies the possibilities of continuous learning. Peirce used the three categories (Firstness, Secondness and Thirdness) and the triadic relation (Representamen, Object and Interpretant) to create a typology with three trichotomies, as shown in Table 1.

**Table 1:** Peircean typology of signs, based on Noth (1995, p. 45)

| Trichotomy<br>Category | of the representamen | of relation to object | of relation to interpretant |
|---|---|---|---|
| *Firstness* | Qualisign (Tone) | Icon | Rheme |
| *Secondness* | Sinsign (Token) | Index | Dicent |
| *Thirdness* | Legisign (Type) | Symbol | Argument |

With this typology, Peirce had conceived 10 possible classes of signs: With this typology, Peirce had conceived 10 possible classes of signs: 1. (Rhematic Iconic) Qualisign,[2] e.g. a non-specific feeling of red; 2. (Rhematic) Iconic Sinsign, e.g. a particular drawing of an ox, carved in a stone, in a cave, recognized by its similarity to an ox; 3. Rhematic Indexical Sinsign, e.g. a subtle non-identified cry, attracting our attention to the person crying; 4. Dicent (Indexical) Sinsign, e.g. a weathercock, as affirming the wind direction here and now; 5. (Rhematic) Iconic Legisign, e.g. a generic diagram, apart from its factual individuality;[3] 6. Rhematic Indexical Legisign, e.g. the general idea behind a demonstrative pronoun like "this," or "that;"[4] 7. Dicent Indexical Legisign, e.g. a recognized traffic sign, planted in the ground, specifying that a particular traffic law is applied there;[5] 8. Rhematic Symbolic (Legisign), e.g. a common noun; 9. Dicent (Symbolic Legisign), e.g. an ordinary proposition; and 10. Argument (Symbolic Legisign), e.g. a syllogism.[6]

Following the unbalanced tripod representation introduced here, and considering that each element in the tripod can be either a possibility, an existent or a law, Figure 3 shows the diagrams corresponding to each one of the ten classes of signs proposed by Peirce. White circles in dashed lines means that the element is a mere possibility (Firstness), while grey circles in continuous line correspond to elements that are true existents (Secondness), and black circles in continuous line correspond to elements that are considered as laws (Thirdness).

In addition to the classification of the signs, some important relations between them must be considered, mostly, the relations concerning *composition* and *government*. *Composition* means that a more complex sign might incorporate other less complex signs. In other words, if a sign might be *broken* into parts, these parts *compose* the sign as a whole, e.g., for the Dicent Indexical Sinsign, Peirce says that "[...] is any object of direct experience, in so far as it is a sign, and, as such, affords information concerning its Object. [...] Such a sign must involve an Iconic Sinsign to embody the information

---

[2] Parenthesis indicates that some terms can be omitted due to redundancy, e.g, a Qualisign must be Rhematic and Iconic, so it can be referred simply as a Qualisign.

[3] This is the case every time an Iconic Sinsign is recognized as an instance of a more general law ruling every single instance of it. The Iconic Legisign is the interpretant of the Iconic Sinsign.

[4] Do not confound that with the words "this" or "that", which are Rhematic Symbolic Legisigns. In a technical sense, the Rhematic Indexical Legisign is the interpretant of a Rhematic Symbolic Legisign, after its interpretation. In fact, when we are reading a text and the word "this" appears, it is first an Iconic Sinsign (the ink at the paper), which is then interpreted as an Iconic Legisign (a recognized word), which is further interpreted as a Rhematic Symbolic Legisign, which is finally interpreted as a Rhematic Indexical Legisign.

[5] The traffic sign, in itself is just a Rhematic Symbolic Legisign, interpreted in the sense of a traffic law. What constitutes the Dicent Indexical Legisign is the fact that this traffic sign is planted in the ground, providing an affirmation that the traffic law, represented by the Traffic Sign (just a Rheme composing the Dicent) is ruling at this location.

[6] Examples adapted from (Noth, 1995, p. 45).

and a Rhematic Indexical Sinsign to indicate the object to which the information refers." (CP 2.257). So, every Dicent might be decomposed into a particular set of Rhemes (fragments of the Dicent), which, as a whole, compose it. And *government* means that all Legisigns (Types) exist through their Sinsigns (Tokens or Replicas), e.g., Peirce claims that

> "An Iconic Legisign (e.g., a diagram, apart from its factual individuality) is any general law or type, in so far as it requires each instance of it to embody a definite quality which renders it fit to call up in the mind the idea of a like object. [...] Being a Legisign, its mode of being is that of governing single replicas, each of which will be an Iconic Sinsign of a peculiar kind." (CP 2.258)

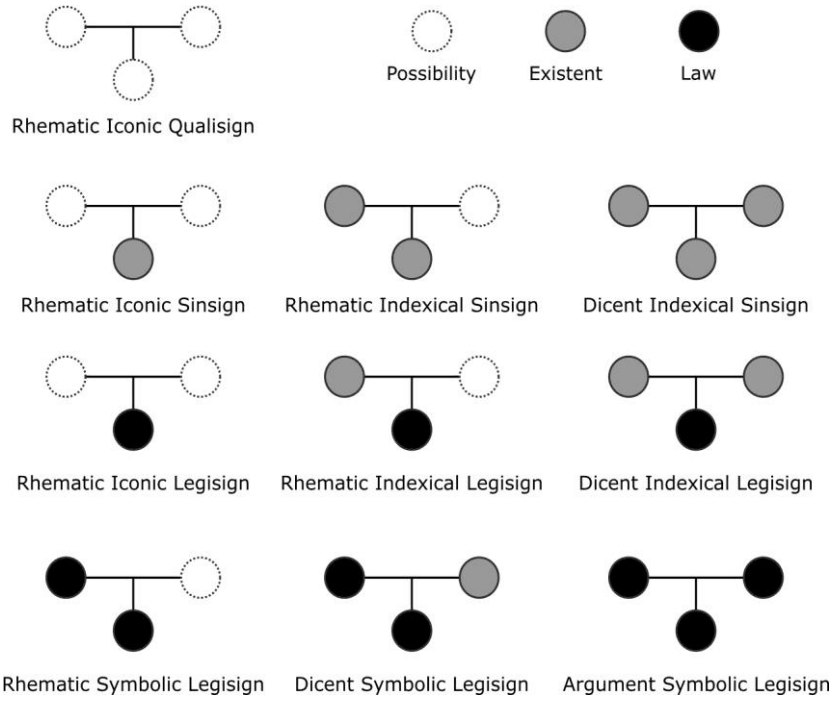



**Figure 3:** The ten classes of Peircean semiotics according to the proposed diagrammatic representation

Taking into account these relations, Figure 4 aims to represent all possibilities for the ten classes of signs. In this figure we adopted the following convention:

- Continuous arrows stand for *composition* and dashed arrows stand for *government;*
- Legisigns are divided into three types, embracing: LI (Laws of Instantiation), LC (Laws of Coding), and LR (Laws of Reasoning);

— In the *composition* scheme, black arrows (a) mean ordinary composition, purple arrows (b) represent the arbitrary associative laws of coding and orange arrows (c) represent agglutination of sentences to form an argument;
— In the *government* scheme, blue arrows (1) represent genuine necessities, green arrows (2) represent arbitrary necessities and the red arrow (3) represents meta-necessities (laws of laws).

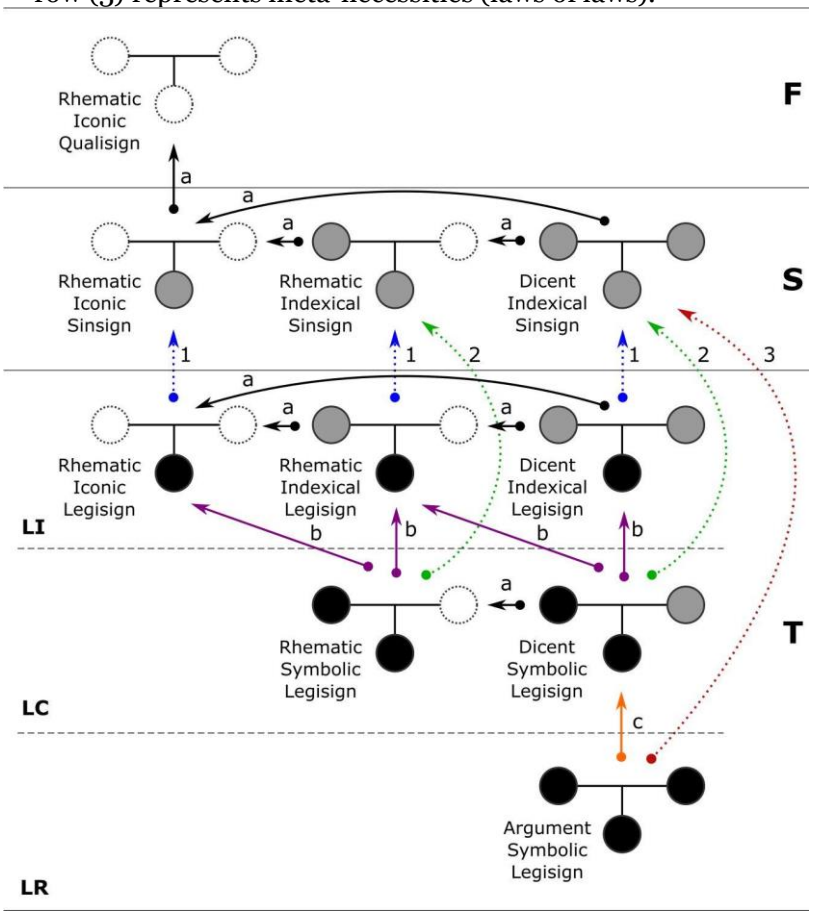



**Figure 4:** The ten classes of signs and their relations of composition and government (based on Camargo, 2018, p. 45)

Composition and government relations will be resumed in the next sections where LI, LC and LR will be clarified. For now, it is necessary to introduce the notion of semiosis as a process. This means that in the space-time framework, a sign can be interpreted into another sign called its interpretant, and this interpretant, for its time, keeps the same potential of interpretability (to be interpreted into a further sign), making the process continuously going on in the direction of the Final Interpretant. This direction

is merely a tendency as the Final Interpretant cannot be really reached because, if so, it would mean that the absolute truth concerning the object was obtained, which is impossible. Even so, as signs grow, the knowledge about the object increases more and more (see Figure 5). Peirce said about the Final Interpretant that:

> "... We must also note that there is certainly a third kind of Interpretant, which I call the Final Interpretant, because it is that which would finally be decided to be the true interpretation if consideration of the matter were carried so far that an ultimate opinion were reached." (CP 8.184)

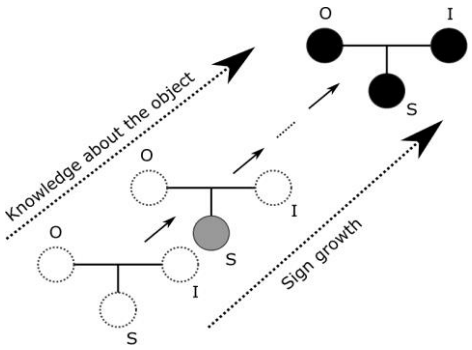



**Figure 5:** Representation of the semiotic process (semiosis)

Finally, it is important to understand the dynamics by which a Dynamical Object, present at the environment, becomes known by an interpreter's mind (see Figure 6). The interpreter does not have direct access to it. Instead, this access is always mediated by a sign. But a single sign cannot bring a full acquaintance of the Dynamical Object. It can only bring a partial facet, a particular aspect of this object, which is called the Immediate Object, the parcel of the Dynamical Object, which is conveyed by a single sign. This particular aspect of the Dynamical Object, is consolidated internally in the interpreter's mind by the generation of the Immediate Interpretant, during a semiosis process. As soon as many different signs bring different aspects of the same Dynamical Object, they are integrated into the Dynamical Interpretant, which can be viewed as the ongoing, best understanding of the Dynamical Object, as long as the interpreter receives signs related to this Dynamical Object. The Final Interpretant is just an ideal target, which will never be achieved, supposing that an infinite number of signs related to this same Dynamical Object could provide a complete understanding of the Dynamical Object. This process is detailed in next section.

## 4. FROM SIGNALS TO KNOWLEDGE

Considering an *agent* (at this moment, natural or artificial) in relation to its *environment*, we assume here that there is a world outside the agent that cannot be reached by direct access, but only indirectly by means of signs, provided by sensors or internal memories. The agent is equipped with *sensor* devices that can capture and measure certain signals coming from the world or, most precisely, signals that correspond to partial *properties* of the objects of the world that can be possibly sensed. Thus, a *signal* is considered here as the substrate of signs, or as "the vehicle of semiotics" which "is opposed to the sign since it is only its physical embodiment" (Noth, 1995, p. 80). In this sense, a signal is a term of information theory and should not be confused with the signs themselves.

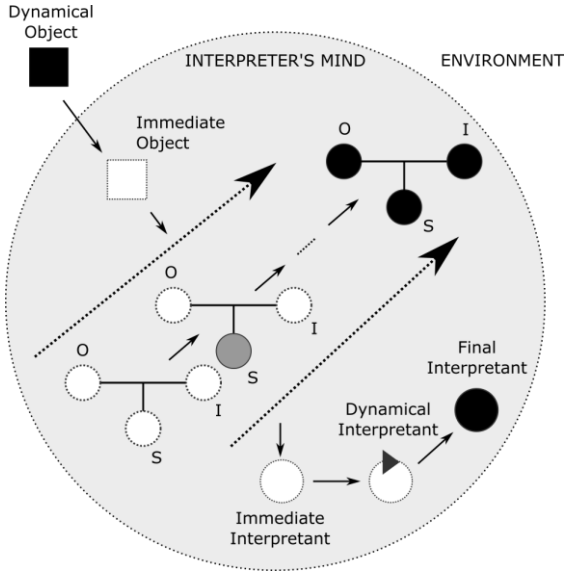



**Figure 6:** The semiotic dynamics between the interpreter's mind and the environment

The environment is populated by things that will be referred to as *presumed existents* or simply existents, which corresponds to the Dynamic Objects, or the real objects that will be represented by a sign.[7] When an agent captures the signals coming from a presumed existent and uses them to determine its actions, it plays the role of an interpreter. The term "presumed" is used to reassure that it is impossible to claim its real state of existence.

[7] In Peirce's words, "That thing which causes a sign as such is called the object (according to the usage of speech, the 'real,' but more accurately, the existent object) represented by the sign: the sign is determined to some species of correspondence with that object." (CP 5.473)

The agent, or interpreter, has a mind that, in the case of artificial agents, corresponds to its control system and memory, which are used to turn signals into signs, transforming the signals information into knowledge (Gudwin, 1999, 2001). Sensors and/or internal memories can be used as sources of signals, by agents, to identify and create compatible models of the world. Besides existents and their properties, things in the real world must actuate following natural regularities or necessities.

Sensors are sources of signals, which should be considered regarding three relevant aspects: *Transduction, Intensity* and *Position/Orientation. Transduction* is a unique capability of sensors, in establishing a natural analogy between different properties of existents with a prototypical property to be used internally by an agent, for processing information within the agent's mind. For each different sensor, a certain kind of property is being measured and translated to this prototypical property, which can be stored and interpreted, e.g. chemical-electric sparks in biological organisms or digital numbers stored in computer memories. *Intensity* corresponds to the magnitude of the signal, compared to a reference physical property, which can be understood as relative numbers that can change as time passes. Finally, *Position/Orientation* has to do with the fact that sensors are space located, so the signal they generate maintains a spatial relation between sensed properties and the agent's own position/orientation. Thus, at each time step, each sensor generates a signal integrating intensity combined to position/orientation, which is then accumulated in a windowed queue, in order to register the passage of time, creating a spatial-temporal dynamics in the manifold of senses. In the interpreter's mind, these signals are conveyed into signs. The most elementary ones are the Qualisigns, which stands in isomorphic relation to the properties of the presumed existents,[8] remembering that this is always a partial process as the presumed existent cannot be captured in its wholeness.

### 4.1. Indexicality and iconicity in sensors

Most Peircean semioticians consider sensors as sources of indexical signs. And there is a reason for that. Peirce himself had written that:

> "For the acceleration of the pulse is a probable symptom of fever and the rise of the mercury in an ordinary thermometer or the bending of the double strip of metal in a metallic thermometer is an indication, or, to use the technical term, is an index, of an increase of atmospheric temperature, which, nevertheless, acts upon it in a purely brute and dyadic way. In these cases, howev-

[8] The terms property and quality can be used to the same signification, but in this paper, the term property is used in reference to the features of the existents of the real world and the term quality is used in reference to the signs

> er, a mental representation of the index is produced, which mental representation is called the immediate object of the sign; and this object does triadically produce the intended, or proper, effect of the sign strictly by means of another mental sign; and that this triadic character of the action is regarded as essential is shown by the fact that if the thermometer is dynamically connected with the heating and cooling apparatus, so as to check either effect, we do not, in ordinary parlance speak of there being any semeiosy, or action of a sign, but, on the contrary, say that there is an 'automatic regulation,' an idea opposed, in our minds, to that of semeiosy. For the proper significate outcome of a sign, I propose the name, the interpretant of the sign." (CP 5.473)

Nevertheless, in this work, we postulate that, regarding sensory processes, there are two possible interpretations about the types of signs sensors can produce. The first interpretation matches the general consensus of indexicality: it is evident that a mercury thermometer acts indicating another object's temperature (the temperature of the air surrounding the thermometer). But what can be told if we are feeling this temperature by ourselves, using the ability of our skin to feel it, for that purpose? It is important to understand that both the mercury thermometer and our skin cells are temperature sensors. In fact, they are both thermometers, maintaining certain isomorphism between the properties of the world and the qualities perceived. The difference is that the mercury thermometer is an external sensor, to which I might have my attention driven, while in the case of my skin, I am directly connected to it. So, sensors might differ depending on the fact of being (or not) a part of the agent. Under this particular perspective, sensors can also be understood as sources of Iconic Signs (Gudwin, 2014). Even when thinking about thermometers and measured temperatures, both instances can be detected. If one looks to a mercury thermometer and starts making conjectures about the intensity of the measured temperature, comparing the size of mercury column to the temperature being felt by their skin, they are involved in an indexical process, but if one touches a hot surface with their own hands, some similar conjectures could be done, but they would have started from a different type of signs, the Iconic Sinsigns.

In a very careful analysis of Peirce's extensive and intricate system of signs classification, Santaella (2020, pp. 293–306) enhanced her earlier concept of the six degrees of iconicity (Santaella, 1996), which lead to the correspondence with three of the ten classes of signs, the Qualisigns, the iconic Sinsigns, and the iconic Legisigns. From her point of view, the iconicity degrees go from Pure Icons (one degree) to Actual Icons (two degrees), and from them to Hypoicons (three degrees).

Following Santaella, the Pure Icon is a quasi-sign, or a sign reduced to a monadic state as it is something merely mental. It's something that does not even become realized as an idea, it stays in the undefined realm of mere

possibility, and she points to Peirce's own words "For in precision of speech, Icons can represent nothing but Forms and Feelings. [...] No pure Icons represent anything but Forms; no pure Forms are represented by anything but Icons" (CP 4.544). However, when considering outward objects presented to one's mind, a dyadic relation is established through perception process, and this change represents the passage from Pure Icons to Actual Icons. Now, it is not a case of merely mental action but a connection between outer and inner worlds that brings into relation the objects of the world and the mind, it is an act of perception. In this scheme, both Pure Icons and Actual Icons are related to Qualisigns (Santaella, 2020).

On the other hand, the Hypoicons are Iconic Signs operating in the level of Secondness and Thirdness (Sinsigns and Legisigins). The Hypoicons can be divided into Images, Diagrams, and Metaphors according to the mode of Firstness they participate in. Images are the signs that participate in simple qualities (Firstness); Diagrams are those signs that represent relations (Secondness); and Metaphors are those signs that represent parallelism in something else (Thirdness) (CP 2.277). Stjernfelt (2007, pp. 293–306) introduces the instances of hypoicons as follows:

> "Images in this restricted, technical meaning of the word are similar to their object due to some simple quality (color, shape, tonality, size ...); diagrams are all similarity-based signs which refer to their object by means of some skeletal analysis of the object into mutually connected parts. The diagram consists of a sketch-like anatomy of its object—as the most ordinary examples one may point to function diagrams, cake diagrams, column diagrams, matrices—but also, cf. below, a much larger set of icon types. Metaphor, finally, is the picture type which refers to its object via the intermediary of a third object."

An important aspect of Hypoicons is their condition of compositionality. An image can be understood as a topological composite of qualities, or a bundle of qualities that, acting together, shapes the objects of the world; a diagram, in its turn, is a composite of relations between different parts of an object; and, finally, a metaphor exposes some kind of parallelism composition represented by a certain type of law that connects objects not by direct affection, but through some kind of idea.

In this context, embodied sensors can be understood as source of Iconic Metaphors, mapping the objects of the world as an analogy of the properties of such objects (Gudwin, 2014). Figure 7 represents the action of a generic sensor functioning as an interface between the world and the interpreter's mind.

Thus, we might consider icons as the most basic bricks of knowledge that can affect an agent and, as pointed by Peirce: "For a pure icon does not draw any distinction between itself and its object. It represents whatever it may

represent, and whatever it is like, it in so far is. It is an affair of suchness only"(CP 5.74). Being so, the evanescent background of Qualisigns must be developed to most complex signs, first to Hypoicons, which represents figures detached from the background, and from them to even more developed signs as indexes and symbols.

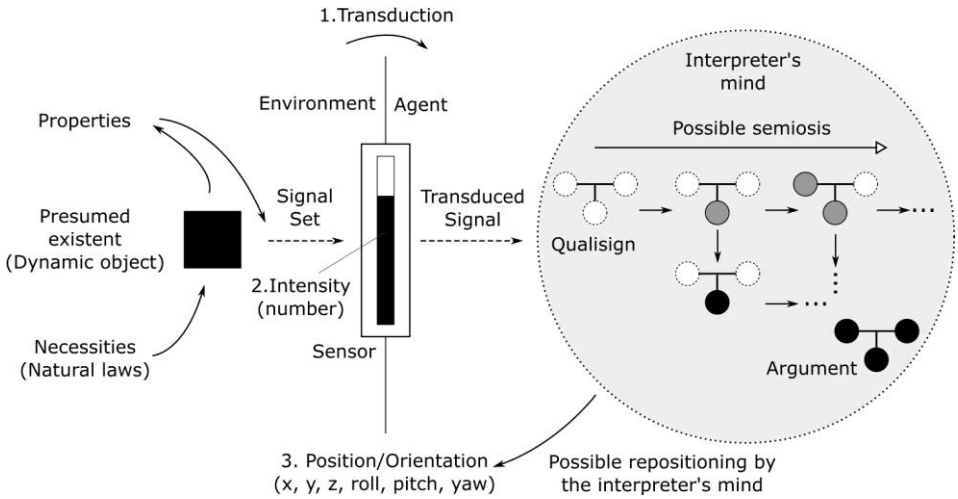



**Figure 7:** The process of signal transdution and its sign representation in the interpreter's mind. Transduced signal occurs on a material substrate according to the nature of the agent (chemical-electric sparks, electric pulse, discrete numbering, etc)

## 4.2. Indexicality and iconicity in sensors

### *4.2.1. Firstness*

Nothing can be claimed for sure about both the existents and the laws ruling their natural interactions. But, due to these laws, existents show some properties and regularities that produce a signal set that, possibly, can be captured by agents equipped with appropriate sensors, e.g., an existent that has the property of reflecting electromagnetic waves in a band between the infrared and the ultraviolet can be perceived by a regular human being. Human eyes have the capacity to transduce photon beams to chemical-electric pulses (transduced signal) that subsidize the agent to create a mental map of the existent. This representation is not a high fidelity copy of the original, but, as written by Vieira (1994, p. 16): "the more an organism can generate environment isomorphic mapping the more it will be near of 'ideal objectivity', and more capable of surviving it will be."

Thus, considering Peircean semiotics, the mapping process generates the Immediate Objects, but, in this first moment, it is just a glimpse of the world, a diffuse background where no detached figure is yet in mind. At this

moment, only Qualisigns affect the agent mind. This is the domain of Firstness, where everything is a mere possibility. It represents the most basic agent–environment interface, and it is the first step of the process toward getting knowledge about the world, e.g., someone stands on an open field with a small house at the left and there is a weathercock on the house's roof, but at this first moment, the eventual interpreter is only able to feel sensory traces of them, through the manifold of senses. Everything is just a diffuse background of evanescent qualities (See Figure 8).

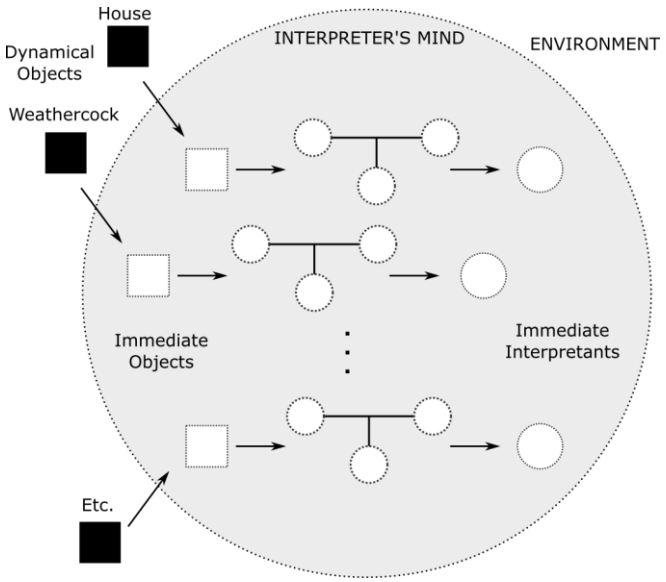



**Figure 8:** Representation of Firstness, when only Qualisigns affect the interpreter's mind, which is surrounded by a diffuse background of evanescent qualities that generate only Immediate Interpretants as semantic potentialities

### *4.2.2. Secondness*

If Firstness corresponds to mere possibilities, the domain of Secondness is where existents became actualized as figures detached from the background. Now, there are three classes of sinsigns affecting the interpreter's mind: the Rhematic Iconic Sinsign, the Rhematic Indexical Sinsign and the Dicent In- dexical Sinsign. The first one corresponds to the impressions of an existent (a thing is perceived), it is a Sinsign because it is, in itself, a part of existence, but the Object and the Interpretant remain as mere possibilities. It is an Icon because it is recognizable due to its similarity with its object. And it is a Rheme, because it is only a part of a possible proposition; the second corresponds to the apprehension of the existent and to the relations that this existent maintains with other existents. It is an Index because it drives attention to the existent it indicates, but the Interpretant remains as

a mere possibility; and the third, finally, turns the Interpretant into a fact, or a judgement (about something), it is the expression of the existent's predicates, a composition that leads to an episode of knowledge.

Then, considering the process of impression-apprehension-expression in the example of a distracted person standing on an open field, suddenly, a gentle breeze moves the arrow of the weathercock and a brief high-pitched metallic sound reaches the agent's ears. The sound makes something to be detached from the background. It is just the primary manifestation of an existent represented by Rhematic Iconic Sinsigns (impression). On a second moment, the attention of the interpreter turns his eyes to the weathercock, Qualisigns and Rhematic Iconic Sinsigns still actuate but now they are involved by Rhematic Indexical Sinsigns that maintain the semiotic process going on: the sound points to the weathercock, which points to the wind (apprehension). Now, the interpreter's mind are populated by Dicent Indexical Sinsigns, which allow the mind to start making judgements about the wind, about its speed and direction, possibly, its temperature could indicate a change in the weather conditions and the necessity to find a shelter, etc (expression). Semiosis keeps going on as new Sinsigns affect the agent, working in the composition of an episode of knowledge (See Figure 9).

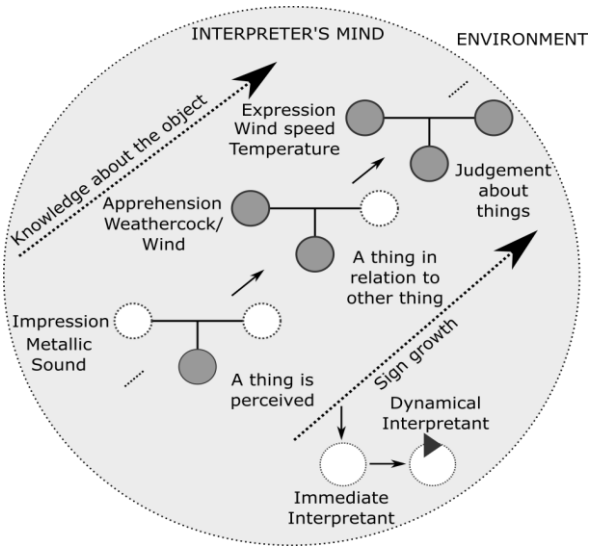



**Figure 9:** Representation of the Secondness, a perceived sound makes a figure to detaches from the background (the weathercock), which leads the attention of the interpreter to make judgements about the weather. At this moment Immediate Interpretants persist but they are accompanied by Dynamical Interpretants as the direct effect produced by a sign

However, if all fact that takes place in the Secondness just represent unprecedented events with no possibility of new occurrences in the future, no regularity, or law, would be recognized, and, consequently, nothing would

be converted to knowledge. In this sense, each Sinsign must be a replica (or a Token) of a Legisign. "The Replica is a Sinsign. Thus, every Legisign requires Sinsigns. But these are not ordinary Sinsigns, such as are peculiar occurrences that are regarded as significant. Nor would the replica be significant if it were not for the law which renders it so" (CP 2.246). Therefore, judgements are possible only due to the actuation of another category: Thirdness.

### *4.2.3. Thirdness*

Recurrent events actualized in Secondness are specific instances of general laws. Sinsigns (Tokens) are governed by Legisigns (Types). There are six classes of Legisigns, and they will be introduced here divided into three subclasses in accordance with the types of laws they represent: Laws of Instantiation (LI), Laws of Coding (LC) and Laws of Reasoning (LR).

Laws of Instantiation are represented by Rhematic Iconic Legisigns, Rhematic Indexical Legisigns and Dicent Indexical Legisigns, each of them corresponding respectively to the types of their tokens: Rhematic Iconic Sinsigns, Rhematic Indexical Sinsigns and Dicent Indexical Sinsigns. These laws represent one dimension of what Hoffmeyer calls code-duality, the dimension of the continuous:

> "As analog [continuous] codifications, organisms recognize and interact with each other in ecological space, whereas as digital [discrete] codifications (genome), they are passively carried forward in time from generation to generation [...]. Seeing from this perspective, life must be understood as *semiotic survival*—survival via a fundamental code-duality" (Hoffmeyer, 2008, p. 80)

If the continuous dimension is represented by the Laws of Instantiation, the discrete dimension of code-duality—which Hoffmeyer exemplifies through genome—is represented by the Laws of Coding: Rhematic Symbolic Legisigns and Dicent Symbolic Legisign. The rhematic character of the first one indicates one unique piece of information as a word, and the dicent character of the second demonstrates the connection of particular pieces of information to a predicate as a sentence. The Laws of Coding can be taken as the elements of a table, where two columns contain, each one, two different sets of terms—set $T_1$ containing the terms $t_{11}$, $t_{12}$,... $t_{1n}$, that correspond to words; and $T_2$ containing the terms $t_{21}$, $t_{22}$,... $t_{2n}$ that correspond to predicates, and each line represents the imposed, or arbitrary, relation between these terms ($t_{11}$ points to $t_{21}$, $t_{12}$ points to $t_{22}$,..., $t_{1n}$ points to $t_{2n}$). Finally, the Laws of Reasoning, Argument Symbolic Legisigns, are responsible to mediating the associative relations between continuous and discrete codes, acting as a set of meta-laws to consolidate knowledge and to allow learning of new habits (See Figure 10).

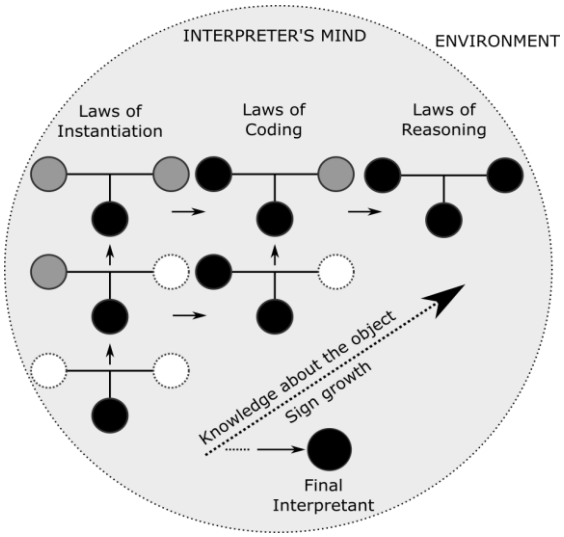



**Figure 10:** Representation of the Thirdness, and its three types of laws: Laws of Instantiation (continuous, classification), Laws of Coding (discrete, consolidation) and Laws of Reasoning (continuous/discrete, learning, laws of laws). It is the moment for the presence of the Final Interpretants that work as a tendency that will not be achieved but promote knowledge

## 5. THE WHOLE PROCESS OF KNOWLEDGE ACQUISITION THROUGH SIGNS: FROM THE THREE WORLDS OF POPPER TO THE WORLDS OF IDEAS AND THE FRAGMENTS OF REALITY

This section introduces some preliminary thoughts about reality based on possible relations between Popper and Peirce. It presents a digression that would help future development in Artificial Intelligence. Despite its speculative character, we consider it relevant to introduce the theme in this paper.

### 5.1. The three worlds of Popper

Trying to represent reality, Karl Popper introduced the concept of three worlds, which were called world 1, world 2, and world 3. The original idea of Popper was not to establish three independent parts of reality, but three levels of it that interact and affect each other. In his own words:

> "There is, first, the world that consists of physical bodies: of stones and of stars; of plants and of animals; but also of radiation, and of other forms of physical energy. I will call this physical world 'world 1'. [...] There is, secondly, the mental or psychological world, the world of our feelings of pain and of pleasure, of our thoughts, of our decisions, of our perceptions and our observations; in other words, the world of mental or psychological states or processes, or of subjective experiences. I will call it 'world 2'.[...] By world 3 I

> mean the world of the products of the human mind, such as languages; tales and stories and religious myths; scientific conjectures or theories, and mathematical constructions; songs and symphonies; paintings and sculptures. But also aeroplanes and airports and other feats of engineering." (Popper, 1972, p. 143)

Popper divides world 1 into the world of non-living physical objects and the world of biological objects, claims that world 2 could be subdivided in various ways, e.g. into conscious experiences and dreams, and points that from world 3 many possible sub-worlds can be distinguished, e.g. the world of science from the world of fiction; and the world of music and the world of art from the world of engineering.

If the idea of knowledge is conceived as the set of representations that can map reality to the agent's mind, and this knowledge is the element that allows cognition to perceive the world and transform it in the benefit of the agent, then the three worlds of Popper can be understood as the things that can be the objects of cognition and also everything that can be known, being the world 1 as the world of knowable things, world 2 the world of the things already mapped inside the human mind, and the world 3 as the result of the human cognition that affects back the world 1.

The three worlds of Popper, in some sense, correlate to the categories of Firstness, Secondness, and Thirdness of Peircean semiotics since world 1 points to the objects of the reality that can be possibly perceived, world 2 to things that take place inside the human mind, and world 3 to things that can be created by humankind and creation presupposes the use of laws to compound complex objects. But, the similarity ceases immediately as Popper considers these worlds in a pluralistic scheme:

> "What have I as a pluralist to say to the materialist monist and to the dualist? First of all, I am, like the dualist, prepared to agree with much that the materialist monist says; in fact, with everything except his denial of a world 2 of experiences and of a world 3 of abstract objects such as the Fifth Symphony. And similarly, I agree with all that the dualist says, except with his implicit belief that the Fifth Symphony is to be identified with our experiences of hearing it, or of remembering it." (Popper, 1972, p. 148)

Peirce's understanding of reality, instead, claims absolute continuity. He says that "yet, the reality of continuity once admitted, reasons are there, divers reasons, some positive, others only formal, yet not contemptible, for admitting the continuity of all things" (CP1.170). And claims that "now the doctrine of continuity is that all things so swim in continua" (CP 1.171). So, considering Popper's effort to divide reality into worlds, but trying to complement this approach with Peirce's categories in continuous relations, next subsection will propose another division of reality into worlds.

### 5.2. The Worlds of Ideas and Fragments of Reality

All systems of categories are attempts to classify the kinds of elements that appear in the mind. These elements can be addressed, individually, as ideas, and elementary ideas can be linked together in order to generate more developed ones. Considering Peircean semiotics, signs are the elements that allow ideas to appear in the mind, and semiosis is the process that relates one idea to another. Popper's world 1 (and possibly part of world 3) can be viewed as the world around us, our environment. Both world 2 and (part of) world 3 can be viewed as worlds related to things appearing in our mind. World 2 include the elements in our mind that can be put in correspondence with elements in world 1. World 3 include the creations of our mind that extrapolate world 1. In this work, we propose a different (but somewhat related) conceptualization, inspired in Peircean ideas. We focus first on what we might refer to the three Worlds of Ideas, categorizing the three different kinds of ideas that can appear in our mind.

These are the World of Possibilities (related to the category of Firstness), the World of Existence (related to the category of Secondness) and the World of Laws (related to the category of Thirdness). Within each of these worlds, we also have embedded three other worlds that are related to three different levels in which we can describe the reality around us (Popper's world 1). These are the World of Senses, the World of Things, and the World of Scenes. Figure 11 aims to represent this intricate organization.

The World of Existence is somewhat equivalent to Popper's world 1. In fact, it is not exactly world 1, but the ideas in our mind regarding world 1. It includes everything we believe is really happening around us, our best guess of what is real, things we may repute as real facts, what we might call "the truth". Feelings of this world are feelings we really felt, or feelings we are currently feeling. Things of this world are things we believe are really there, things we believe are real, in the past or in the present. Scenes of this world are scenes we experienced by ourselves, or which by the testimony of others we believe they really happened, or are happening right now. In summary, the World of Existence includes our naive understanding of what we might call reality, or the real world, but which might better be called actuality or existence. We decided to start our description of the three Worlds of Ideas by the World of Existence, because usually this is the point of contact of our mind with reality. But, as pointed out by Peirce, the reality is more than the World of Existence. It includes also the World of Possibility and the World of Laws. In fact, according to Peirce, the World of Existence is related to Secondness.

Now let's investigate the other two worlds of ideas. The World of Possibility (domain of Firstness) is the world including everything that is possible, or at least we believe is possible, in the world of existence. It includes imagination, speculation, hypothesis, plans for the future, exploration of

scenarios, etc. The World of Possibility is the world where fiction and imagination take place. It is the world that works as the playground of the mind, where we situate things we know are not part of existence, but (we speculate) might be. Feelings in this world are just possible feelings, not those that really happened, or are happening. We can think of the World of Possibility as either atemporal, out of time, or as in the future, something that might possibly happens in the future. Things in the World of Possibility are just imaginary things, things that might exist, but without any commitment with things that really exist. Scenes in the world of possibility can be the recreation of scenes that really happened at the world of existence, but without any implication that they really happened. The World of Possibility is where we situate our interpretations, when we read a book of fiction, when someone tells us an invented story. We know they are not real, even though, they might be. While interpreting text, the World of Possibility is where we situate things when we talk about "a horse," or "a red apple." It is not a particular horse we get in touch during our life, but just a possible horse. It is not a particular apple we found, or might find, nor with a particular kind of red, but just any possible red.

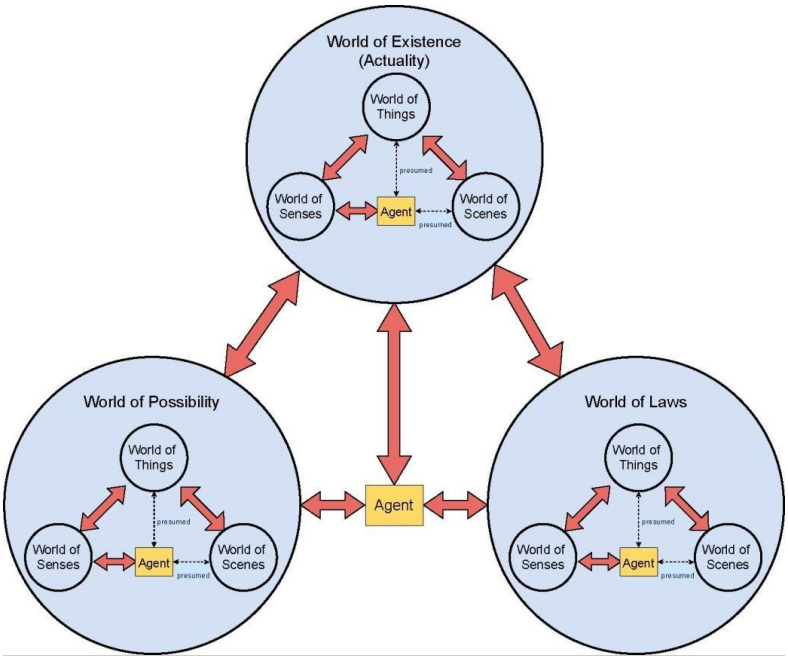



**Figure 11:** The worlds of ideas and the fragments of reality

Sane persons, while thinking, are fully capable of making the distinction of what pertains to the World of Existence, and what pertains to the World of Possibility. Illusions are things we erroneously situate in the World of

Existence, but are actually part of the World of Possibilities. Also, people with mental problems might make a confusion between these two worlds. In fact, anyone might be subject to this mistake, in some situations. Nevertheless, in its process of interpreting signs, an agent should be always classifying things as being either a part of the World of Existence, or of the World of Possibility. Besides that, everything in the World of the Existence might have a dual in the World of Possibility, as everything that really is, must first, be possible.

Finally, the last of the three Worlds of Ideas is the World of Laws (domain of Thirdness). There reside all the ideas regarding concepts, categories, types, habits of behavior, patterns, learning algorithms, rules, etc. Even though the concept of law is something that we can intuitively understand, at the same time it is quite difficult to precisely define. A law is a generic term we use to abstract an idea that mediates the relation between other ideas. A law is the reason by which many other different ideas are bounded together in becoming instances of it (the law). In set theory, it is the membership rule that defines if an element is (or is not) a member of a set. In physics, laws capture the generality of phenomena and allow them to be described in terms of things that keep repeating themselves in different situations. In human social interactions, laws are human arbitrations for enforcing certain habits of conduct among a group of people. The concept of law embraces the idea of generality, of an implicit reason for putting together things that might have their own individuality, but at the same time share some kind of commonality. Thus, in the World of Laws we locate all the concepts governing all other kinds of ideas.

These concepts govern both the ideas pertaining the world of possibility and the World of Existence. These are meant to be instances of these concepts. So the feelings in the World of Laws are all the categories of feelings (or senses) that are related to the feelings in the World of Possibilities and the feelings in the World of Existence. The things in the World of Laws are all the categories of things that are related to the things in the World of Possibilities and the things in the World of Existence. And the scenes in the World of Laws are all the categories of scenes that are related to the scenes in the World of Possibilities and the scenes in the World of Existence.

Given that, we can see that the same three Worlds of Senses, Worlds of Things and World of Scenes do appear in the World of Possibilities, the World of Existence and the World of Laws. But the ideas there, besides sharing some commonality, are not the same. We refer to these three worlds as the Fragments of Reality, in three different contexts. In terms of existence, the existence can be fragmented (segmented) into scenes, involving different things, where different (sensed) properties might change (or remain constant), along time passes.

Now, permeating all the three Worlds of Ideas (and their Fragments of Reality), we have signs, representing the different aspects of reality. The identification often classes of signs, separated into three categories, presupposes that all types of ideas can be encoded to express these different aspects of reality.

In a natural or artificial agent, all these signs are supposed to manifest themselves either on signals coming from sensors and actuators, or in signals stored into internal memories. In natural agents, these signals might be chemical/electrical signals located in specific body cells, like e.g. neurons, or muscular cells. In artificial agents, we might generalize that all the signals are stored into computer memories, given that some memory addresses are in fact mapping into the agent's sensors/actuators. Through its sensors and actuators, the agent can only reach the World of Senses, leaving the World of Things and the World of Scenes to be always presumed, even in the World of Existence and in the World of Possibilities. Also, not all of the possible/existing properties can be captured by the agent. So, depending on the nature of the agent and the capacity of its sensors, different properties might be sensed/represented. The property in itself will depend on the sensor/actuator where the signal comes from.

In this scheme, the signals coming from the agent's sensors represent different features of the Dynamical Objects supposed to exist in the reality. These dynamical objects are the fragments of reality being represented by signs. They might be senses, things or scenes, depending on the signs representing them. The most basic kind of sign, the *Qualisign*, is only capable of representing a sense in the World of Possibility. It is used to represent just a hypothetical sense (an imaginary or generic one). In order to represent a sense that was really sensed, by an agent, a sense at the World of Existence, we might require a *Rhematic Iconic Sinsign*, governed by a *Rhematic Iconic Legisign* (which represents a law at the World of Laws). When an agent captures, in a given instant of time, a particular measuring from a particular sensor, this causes the creation of a Rhematic Iconic Sinsign at the agent's internal memory. As soon as this Rhematic Iconic Sinsign is recognized as an instance of a Rhematic Iconic Legisign, the dynamics of the World of Senses is completed. Despite the status of Secondness of the Rhematic Iconic Sinsigns, their presence as mere possibilities, as becoming from a hypothetical Thing, from the World of Possibilities, denotes their iconic condition.

Things can be represented as bundles of properties. So, in order to represent things, either from the World of Possibilities or from the World of Existence, we need to use *Rhematic Indexical Sinsigns* governed by *Rhe matic Indexical Legisigns.* They are indexes, because they do not have in themselves the properties, but they point to icons that represent these properties. So, each Rhematic Indexical Sinsign will be pointing to multiple Rhematic

Iconic Sinsigns, if this thing is from the World of Existence, or Qual- isigns, if it represents a thing at the World of Possibilities. So, in fact, a Rhematic Indexical Sinsign is a bundle of Rhematic Iconic Sinsigns, representing multiple senses at the World of Existence, or a bundle of Qualisigns, representing multiple senses at the World of Possibilities. Rhematic Indexical Legisigns are laws, from the World of Laws, used to represent classes of things.

Finally, the agent can make judgements about the things, their properties and how these properties change over time forming scenes. In order to represent these scenes and other kinds of judgment about the particular state of any particular thing being a part of a scene, we might use *Dicent Indexical Sinsigns* governed by *Dicent Indexical Legisigns*. These are indexes because they point either to Rhematic Indexical Sinsigns representing the things participating at the scene, or to Rhematic Iconic Sinsigns or to Qualisigns representing a particular sense, from the World of Possibilities or the World of Existence, while being used or not to characterize the scene. This scene might be from the World of Possibilities, while being a hypothetical (generic or imaginary) scene, or from the World of Existence, while being a scene that really happened. Dicent Indexical Legisigns are laws representing the many classes of scenes sharing some kind of commonality. While perceiving a scene evolving in time, the agent creates a Dicent Indexical Sinsign. As soon as this Dicent Indexical Sinsign is recognized as being an instance of a Dicent Indexical Legisign, the class of scene just happening is recognized. This dynamics makes the World of Scenes.

The three worlds representing the Fragments of Reality: the World of Senses, the World of Things, and the World of Scenes are not confined to a specific World of Ideas. They can be perceived inside the World of Possibilities, the World of Existence, and the World of Laws. What differentiate the Fragments of Reality inside each World of Senses and the World of Things is the combined status of all necessary sensed qualities needed to generate the representation of specific things, and then how these things are involved in the representation of scenes. For example, the idea of a horse involves a bunch of qualities, like form, color, being located in a specific space-time, etc. When all (or at least some of) these qualities are undefined or vague, the idea represents a horse in the World of Possibilities. This might be a generic indefinite horse, or maybe a fictional horse. When all these qualities are defined, the idea represents an existent horse, or a specific horse in the World of Existence.

*Rhematic Iconic Legisigns*, *Rhematic Indexical Legisigns* and *Dicent Indexical Legisigns* are all cases of Laws of Instantiation. These laws are used to represent classes or types, governing instances of different fragments of reality. Beyond the Laws of Instantiation, there are two other classes of signs, the *Rhematic Symbolic Legisigns*, and the *Dicent Symbolic Legisigns* that are cases of Laws of Coding. They govern their Replicas in

such a way that the first represents symbols to other things around (like a word or a group of words not forming a complete sentence), and the second represents symbols of judgments about scenes, things and senses (like a complete sentence). The first governs Rhematic Indexical Sinsigns, and the second governs Dicent Indexical Sinsign, which makes to place them, respectively, in the World of Things and in the World of Scenes. Finally, rests the *Arguments* (Argument Symbolic Legisigns), which are learning rules (Laws of Reasoning) that govern Replicas of Dicent Indexical Sinsigns, which corresponds to the Worlds of Scenes.

The result of the division of reality into worlds based on the theory of signs of Peirce, even being a bit intricate, introduces a new approach to study knowledge acquisition in a continuous framework, which can direct the researchers' attention to once hidden pieces of evidences to conceive better models of the mind.

## 6. FROM KNOWLEDGE TO ACTION

### 6.1. Energetic Interpretants, Actions and Creativity

In last section, two levels of knowledge, based on Peircean Semiotics, were considered: one driving the acquisition and establishment of knowledge, mediated by Laws of Instantiation and Laws of Coding, and one allowing learning, mediated by Laws of Reasoning, leading to habit changes. Now it is time to address how semiosis can be related to actions.

Semiosis is the process by which a sign causes an effect, its interpretant. According to Peirce's earlier ideas, this interpretant was supposed to necessarily be another sign, generating a scheme of infinite semiosis, in which an interpretant is also a sign and, being so, another interpretant should be present in an unending process. But Short (2004) points out that, after 1904, Peirce expanded his original point of view, proposing that an interpretant need not always be another sign. In this expanded comprehension, even though the genuine effect of a sign is to generate another sign (a thought-sign), degenerate cases of interpretants might be actions or feelings as well. Peirce's claims that:

> "... Taking sign in its broadest sense, its interpretant is not necessarily a sign. [...] We may take a sign in so broad a sense that the interpretant of it is not a thought, but an action or experience, or we may even so enlarge the meaning of sign that its interpretant is a mere quality of feeling. A Third is something which brings a First into relation to a Second. A sign is a sort of Third. How shall we characterize it? Shall we say that a Sign brings a Second, its Object, into cognitive relation to a Third? That a Sign brings a Second into the same relation to a first in which it stands itself to that First? [...] A sign therefore is an object which is in relation to its object on the one hand and to

> an interpretant on the other, in such a way as to bring the interpretant into a relation to the object, corresponding to its own relation to the object. I might say 'similar to its own' for a correspondence consists in a similarity; but perhaps correspondence is narrower." (CP 8.332)

Following Short (2004) explanation, Peirce established a new classification of the interpretants, considering that feelings are monadic, actions are dyadic, and signs are triadic, and, in 1907, Peirce called these types of interpretants emotional, energetic, and logical. In Peirce's words:

> "This 'emotional interpretant,' as I call it, may amount to much more than that feeling of recognition; and in some cases, it is the only proper significate effect that the sign produces. Thus, the performance of a piece of concerted music is a sign. It conveys, and is intended to convey, the composer's musical ideas; but these usually consist merely in a series of feelings. If a sign produces any further proper significate effect, it will do so through the mediation of the emotional interpretant, and such further effect will always involve an effort. I call it the energetic interpretant. The effort may be a muscular one, as it is in the case of the command to ground arms; but it is much more usually an exertion upon the Inner World, a mental effort. [...] In advance of ascertaining the nature of this effect, it will be convenient to adopt a designation for it, and I will call it the logical interpretant, without as yet determining whether this term shall extend to anything beside the meaning of a general concept, though certainly closely related to that, or not." (CP 5.475-6)

So, following this new trichotomy introduced in 1907, we can assume that every action is the energetic interpretant of a previous sign, processed by an agent in its behavioral process. We can split the process of action into three different moments or stages. In the first moment, an action is *proposed*. In a second moment, an action might *be selected*, among many possible actions, which might have been first proposed. Finally, the selected action is then *performed*.

Considering the different means by which an action might be proposed, and applying Firstness, Secondness and Thirdness, we might reach three different kinds of actions:

- Actions of Firstness are spontaneous kinds of actions, usually embedded with some sort of randomness with an exploratory disposition. This kind of actions induces strategies of trial and error that lead to start finding solutions to unknown situations. The main characteristic of an Action of Firstness is that its proposal (or determination) is completely independent of the present situation.
- Actions of Secondness are reactive actions triggered by a sensor or an internal signal. They are reactions to either external or internal stimuli, usually the fruit of a habit of conduct, triggered only due to the per-

ception of an immediate present. So, its proposal (or determination) is a function of the present situation.

— Actions of Thirdness are motivated actions, which are the most complex of them, requiring both the perception of the immediate present and an expected desired future state, which the agent is supposed to achieve. In order to perform Actions of Thirdness, an agent needs to have at least some model of the effect of possible actions when applied to the present state, and the new state to be achieved if these actions are applied. Then, a path of actions (a plan) might be conceived, moving the present situation to the future desired state. So, its proposal (or determination) is a function both of the present situation, and the desired future state the agent is supposed to reach.

We assume that multiple processes of semiosis happening inside an agent's mind might propose a whole set of actions, which will compete to each other in order to be selected, such that the chosen one, in any particular instance of time, will then be performed, or executed. Actions of Firstness will be proposed every time a certain level of doubt is achieved, or the agent is engaged into some sort of exploratory behavior, particularly when the current situation is unclear, or the agent is not certain in what to do. Actions of Firstness might be important during the process of learning by trial and error, such that the agent can build a behavioral model, and create internal habits of conduct. Actions of Secondness usually are the case when such habits are already defined, and the agent now knows how to act, based on the present situation. Actions of Secondness are very useful when the agent does not have time for thinking, but previous experiences have created a set of rules relating specific situations to specific reactions. In this case, they might be proposed in order to repeat a behavior already performed in the past, which might have been proven useful. Instead of simply trying everything, like with the Actions of Firstness, Actions of Secondness might go straight to the point, regarding the current situation. But even though an agent might achieve interesting behaviors with just Actions of Firstness and Actions of Secondness, only with Actions of Thirdness an agent might really succeed in shaping the environment into a future desired state. The problem is that Actions of Thirdness might be very costly to propose. They require a good model for the effect of different actions, and might require elaborate planning strategies in order to move from the current present situation up to the desired future state. Also, if unexpected changes at the environment do appear, they might require some sort of re-planning in order to be successful.

Different strategies might be used to generate Actions of Firstness, Actions of Secondness and Actions of Thirdness. Boden (1998) introduces a very interesting typology of creativity that can be used to propose new actions. In her words:

> "There are three main types of creativity, involving different ways of generating the novel ideas. The first type involves novel (improbable) combinations of familiar ideas. Let us call this 'combinational' creativity. Examples include much poetic imagery, and also analogy-wherein the two newly associated ideas share some inherent conceptual structure. [...] The second and third types are closely linked, and more similar to each other than either is to the first. They are 'exploratory' and 'transformational' creativity. The former involves the generation of novel ideas by the exploration of structured conceptual spaces. This often results in structures ('ideas') that are not only novel, but unexpected. [...] The latter involves the transformation of some (one or more) dimension of the space, so that new structures can be generated which could not have arisen before. The more fundamental the dimension concerned, and the more powerful the transformation, the more surprising the new ideas will be. These two forms of creativity shade into one another, since exploration of the space can include minimal 'tweaking' of fairly superficial constraints. The distinction between a tweak and a transform is to some extent a matter of judgement, but the more well-defined the space, the clearer this distinction can be."

It is interesting to notice that Boden's typology intrinsically follows Peirce's ideas of Firstness, Secondness and Thirdness. Exploratory creativity is clearly an instance of Firstness, as it requires nothing to create a new one, using only the structure of the conceptual space. Transformational creativity is an instance of Secondness, as it requires a first in order to transform it and create a new one. And Combinational creativity is clearly a Thirdness, as it requires both a first and a second in order to create a new one. Assuming now that the conceptual space is the actuation space, each of the kinds of actions (of Firstness, of Secondness, and of Thirdness) can be created using either exploratory, or transformational, or combinatorial strategies.

Now, after possibly a set of actions was proposed (and among them, there might be Actions of Firstness, Actions of Secondnesss, and Actions of Thirdness), the agent needs to select one of them to be executed. Again, in the selection process, we might have three different strategies for action selection, based on Firstness, Secondness and Thirdness. The simpler selection strategy is a *non-deterministic* selection, a strategy using Firstness. This can be a completely random selection, or follow some statistical distribution, based on a set of priorities related to action selection. The main characteristic of the non-deterministic selection is that, following the same principle used in action proposal, the selection does not use the present state for making the choice, but just some fixed *a priori* set of preferences for doing so. The second selection strategy, now a strategy of Secondness, uses the present situation as parameter for choosing the action. In this case, the selection is some function of the present situation, the reason we are calling it the *deterministic* selection, as it is determined by the present situation. Finally, a strategy of Thirdness is the one that performs the choice using both the

present situation and a desired future state the agent is supposed to reach, which we are calling a *goal-based* selection.

After an action is finally chosen, it can be executed, by applying the chosen action parameters to the actuators.

### 6.2. Back to signals

Now, it's time to close the loop. At first, sensors capture and measure the signals coming from the world, allowing the agent to be aware of partial properties of the Dynamical Objects. The signals are transduced to a specific material substrate according to the nature of the agent, which maps the properties of the world into Qualisigns. According to the doctrine of signs of Charles Sanders Peirce, semiosis occurs through the growth of signs, from the most elementary of them, the Qualisigns, to the most developed, the Arguments. This process is not necessarily an infinite one, but it can also end on actions induced by Energetic Interpretants (connected to certain types of creativity). After a set of actions is proposed, one of them is selected, based on some algorithm of action selection, and finally, the action can be executed, by setting up the actuators with a determined set of signals, which might cause a possible change in the environment. This change upon the world makes the sensors capture new signals, making the whole process of cognition go on and again.

This loop is in accordance with Noe. To him, perceiving is a way of acting and perception is not something that happens to us, or in us, but something we do. And he claims that:

> "To be a perceiver is to understand, implicitly, the effects of movement on sensory stimulation. [...] An object looms larger in the visual field as we approach it, and its profile deforms as we move about it. A sound grows louder as we move nearer to its source. Movements of the hand over the surface of an object give rise to shifting sensations. As perceivers we are masters of this sort of sensorimotor dependence. This mastery shows itself in the thoughtless automaticity with which we move our eyes, head and body in taking what is around us. [...] The central claim of what I call enactive approach is that our ability to perceive not only depends on, but is constituted by, our possession of this sort of sensorimotor knowledge." (Noe, 2004, pp. 1–2)

The sensorimotor approach depends on two complementary instances of the whole process. First, it depends on a feedforward instance that will result in some immediate effect on reality, and, second, it depends on a feedback instance that changes the state of the process based on the effect caused by the first instance. A similar approach is being addressed by Hawkins (2004, 2021) since 2002. He developed a new type of artificial neural network, called Dynamic Sparse Network that is not only inspired by biolog-

ical features but also tries to reproduce the structure of the neocortex, resulting in the HTM technology (Hierarchical Temporal Memory).

The network model proposed by Hawkins is a structure of various cortical columns disposed side by side. These columns have two layers, the input layer, and the output layer. Each column is responsible to receive feedforward sensory input from a specific sensor, e.g., from each finger of a hand in a typical position of grabbing an object (See Figure 12). There is combined information on the input layer composed by the sensory input itself and the location from which this sensory input is collected, sensed by means of proprioceptive collateral sensors. This combined representation allows the capture of features at specific locations at the original object being sensed, providing collateral information, which is essential during sign interpretation. Also, the output layer receives feedforward inputs from the input layer, converging to stable patterns representing the object that the hand has grabbed, such that the object can be mapped in the agent's mind. There are two ways to reach this convergence in the second layer: 1. By integration over time as the sensor moves relative to the object; and 2. By modulatory lateral connections between columns that are simultaneously sensing different locations on the same object. Finally, feedback from the output layer to the input layer allows the input layer to predict what feature will be present after the next movement of the sensor (Hawkins et al., 2017).

Hawkins experiment provides an insightful evidence for what we have already proposed in figure 7: that the location (position/orientation) from which a sensed signal comes from is *fundamental* for a proper representation of the world by means of signs. It is only due to this collateral information that signs might act as indexes, realizing its full-fledged semiotic potential for representing the world.

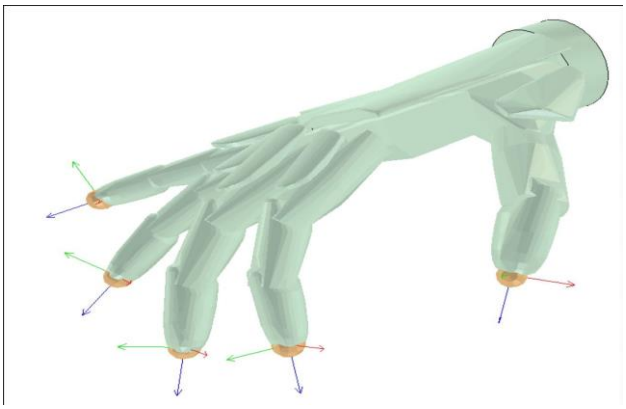


**Figure 12:** Hand in a grabbing position

## 7. SEMIOTICS OF ARTIFICIAL AGENTS

So far the analysis regarded a generic agent. Now we introduce some thoughts about the consequences of a semiotic approach to artificial agents based on digital technology such as a computer or a robot.

Each artificial agent must be programmed in such a way that a proper semiotic process can emerge, in order to build knowledge from data, allowing its use by the agent in the pursuit of its goals and purposes. Semiosis promotes the growth of signs, first from Qualisigns related to the World of Senses, towards Sinsigns and Legisigns related to the World of Things, and later to the World of Scenes. By creating a proper representation for the World of Existence, and abstracting the laws regulating it (forging a representation for the World of Laws), the agent is able to build a representation for the current situation, and with that explore alternative scenarios in the World of Possibilities such that possible courses of action leading to a desired future state might be planned. These dynamics make the system deal with increasingly developed signs, which the most developed of then are the Argument Symbolic Legisign, or simply the Argument.

Following Peirce's categories, an artificial agent is primarily in contact with the signs of Firstness, coming from its sensors, which provide hints about the actual state of the world, what can both be used to drive the agent actions immediately or to evolve more complex signs in memory, to be used in the future. What a sensor senses is different from what the sensor is, and, for an artificial agent, the signal set corresponding to the measurable properties of the world can only be transduced in numbers representing these properties. So, in the inner world of an artificial agent, all qualities of Firstness are numbers, and we assumed here that these numbers are in isomorphic relation to the entities of the world. Moreover, as the properties change in time according to natural laws the set of all measures obtained (numbers in relation to other numbers) also reflect an isomorphic relation, this time with laws, and laws as a generalization can be understood as concepts.

In the context of artificial agents, all concepts can be addressed as data that is used as a substrate for different kinds of signs, as time passes. The process of data flow can be represented by the ultimate techniques of artificial intelligence, such as neural networks, from conventional approaches of Deep Learning to the previous mentioned Dynamic Sparse Networks. No matter which kind of representation is picked, it must have the capacity to computationally model two kinds of laws: 1. The laws that, given a specific representation, can generate replicas of such representation; and 2. Laws that, given a possible replica, verify if it belongs to a general type, and, if so, to which law it corresponds.

So, as isomorphic structures induce the idea of iconicity, these numbers can be taken as icons of the existents, and, for this, the raw data collected by

sensors can be used to create a representational map of the world. But, in Firstness, the set of data is just a "background" of possibilities and some Firstness exploratory actions must be performed to find some coherent pattern that reflects an isomorphic structure to the world. This type of strategy is used in the field of artificial intelligence, mostly when techniques of neural networks and deep learning are used. Obviously, all process of constructing such architectures is mediated by human beings, and terms as "machine learning", and even artificial intelligence can induce the wrong idea of genuine machine intelligence compared to human intelligence. But the essence of Firstness actions allowed by the background of raw data can be addressed to the artificial intelligence field, as it deals with possibility, probability, and, in some sense, with a non-deterministic approach, which is very useful in exploratory behavior. After an artificial intelligence model is developed, its use no longer implies possibilities, but conventional computation with deterministic relations between inputs and outputs, which points to Secondness, the domain of dyadic relations. If some aspects of artificial intelligence can be related to Firstness, and conventional computation to Secondness, Which aspects of the computational field could be addressed to Thirdness? E.g., is it possible to address artificial creativity? Returning to Boden, she claims that:

> "Computer models of creativity include examples of all three types. As yet, those focused on the second (exploratory) type are the most successful.
> That's not to say that exploratory creativity is easy to reproduce. On the contrary, it typically requires considerable domain-expertise and analytic power to define the conceptual space in the first place, and to specify procedures that enable its potential to be explored. But combinational and transformational creativity are even more elusive. The reasons for this, in brief, are the difficulty of approaching the richness of human associative memory, and the difficulty of identifying our values and of expressing them in computational form. The former difficulty bedevils attempts to simulate combinational creativity. The latter difficulty attends efforts directed at any type of creativity, but is especially problematic with respect to the third." (Boden, 1998)

## 8. CONCLUSION

The difficulties pointed by Boden can be found in all attempts to create precise models of the mind that could be reproduced in a computational environment, including the ones inspired by biological entities such as the human brain. Lindsay (2021, pp. 360–364) points that even the most expensive models are not perfect replicas of the object of inspiration. Due to the great complexity involved, the creators of these models need to choose what to include and what must be left outside the model, in other words, what the scientists aim to explain and what they can ignore.

In order to face this challenge, this article tried to expose some principles of the General Theory of Signs developed by Charles Sanders Peirce that can address a new vision about how the minds possibly work. Bringing semiotics to the cognitive science field could be a very fruitful effort in the task of planning and building more efficient artificial agents. To do so, this work proposed a diagrammatic representation of signs, of semiosis, and of the relations between Qualisigns, Sinsigns, and Legisigns that can help researchers to find the essential informational process involved in the sensing-actuating loop that leads to the capacities of getting knowledge about the objects of the world and of actuating back changing these objects.

This interaction starts with the sensor devices that capture and measure signals from the environment, such that these signals must be encoded into signs: first into Qualisigns, and, as the semiotic process goes on, these signs grow in the direction of most developed signs, in order, from Qualisigns to Sinsigns, and from Sinsigns to Legisigns. Thus, semiosis is responsible for increasing the knowledge about the world. Finally, from the knowledge acquired, the agent can act upon the world, changing it and making new signs available. When considering an artificial agent, the implementation of the whole process can be possibly addressed through Dynamic Sparse Networks that conjugate feedforward and feedback treatment of the signals transduced by sensors.

At this moment, the semiotic approach proposed here represents the preliminary efforts in the direction of planning and building Semiotic Artificial Agents. We believe that this is the beginning of a very promising path of research.

## REFERENCES

J. L. Bermudez, *Cognitive Science: An Introduction to the Science of the Mind.* 3rd. Cambridge University Press, Cambridge 2020.

M. A. Boden, *Creativity and artificial intelligence*, In: Artificial Intelligence, 103, 1998, pp. 347–356.

____ , *Minds as Machine: A History of Cognitive Science*, VolS. 1, 2. Oxford University Press, Oxford 2006.

C. E. P. de Camargo, "Semiótica da vida artificial", PhD thesis. São Paulo: Pontifícia Universidade Católica, Faculdade de Ciências Exatas, Curso de Pós-graduação em Tecnologias da Inteligência e Design Digital, 2018.

R. R. Gudwin, *From Semiotics to Computational Semiotics*, *In:* Proceedings of the 9th International Congress of the German Society for Semiotic Studies/ 7th International Congress of the International Association for Semiotic Studies (IASS/AIS), Dresden, Germany, 3–6, 7–11 October, 1999.

____ , *Semiotic Synthesis and Semionic Networks*, In: SEE'01—2nd International Conference on Semiotics, Evolution and Energy, University of Toronto, Toronto, Canada, 10, 2001.

____ , *Peirce and the Engineering Of Intelligent Systems*, In: Death and Anti-Death, vol. 12*:* One Hundred Years after Charles S. Peirce (1839–1914). C. Tandy (ed.), Ria University Press, 2014. Chap. 7, pp. 207–224.

____ , *Computational Semiotics: The Background Infrastructure to New Kinds of Intelligent Systems*, in: APA Newsletters 15-1.1, 2015, pp. 27–38.

J. Hawkins, *On Intelligence: How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines,* Henry Holt & Co, New York 2004.
___ , *Thousand Brains: A New Theory of Intelligence by Jeff HawkinsI,* Basic Books, New York 2021.
J. Hawkins, S. Ahmad, Y. Cui, *A Theory of How Columns in the Neocortex Enable Learning the Structure of the World*, in: Frontiers in Neural Circuits, article 81, 2017.
J. Hessen, *Teoria do conhecimento*, 3rd ed., Martins Fontes, São Paulo 2003.
J. Hoffmeyer, *Biosemiotics: an Examination into the Signs of the Life and the Life of Signs*, University of Scranton Press, Chicago 2008.
G. Lindsay, *Models of the Mind: How Physics, Engineering and Mathematics Have Shaped Our Understanding of the Brain*, Bloomsbury Sigma, London 2021.
A. Noe, *Action in Perception*, MIT Press, 2004.
W. Noth, *Handbook of Semiotics*, Indiana University Press, Bloomington, IN 1995.
___ , *Charles S. Peirce's Theory of Information: a Theory of the Growth of Symbols and of Knowledge*, in: Cybernetics and Human Knowing, 19 (1–2), 2012, pp. 137–161.
C. S. Peirce, *Collected Papers of Charles Sanders Peirce*, Vols. 1–6, C. Hartshorne, P. Weiss (eds.); Vols. 7–8, A. W. Burks (ed.), Harvard University Press, Cambridge, MA 1931-58.
K. Popper, *The Three Worlds*. The Tanner Lecture on Human Values, 1972.
J. Queiroz, *Semiose segundo C. S. Peirce*, EDUC, Sao Paulo 2004.
S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed., Pearson, Hoboken, NJ 2020.
L. Santaella, *Peirce's Broad Concept of Mind*, European Journal for Semiotics Studies, 6, 1994, pp. 399–411.
___ , *From Pure Icon to Metaphor: Six Degrees of Iconicity*, In: Peirce's Doctrine of Signs: Theory, Applications, and Connections, V. M. Colapietro, T. M. Olshewsky (eds.), De Gruyter Mouton 1996, pp. 205–214.
___ , *A teoria geral dos signos: como as coisas significam as coisas*, Cengage, Sao Paulo/SP 2000.
___ , *The Cognitive Function of Iconicity*, ihn: Operationalizing Iconicity, P. Perniss, O. Fischer, C. Ljungberg (eds.), John Benjamins B.V., 2020, pp. 293–306.
T. L. Short, *The Development of Peirce's Theory of Signs*, in: The Cambridge Companion to Peirce, C. Misak (ed.), Cambridge Companions to Philosophy, Cambridge University Press, 2004, pp. 214–240.
F. Stjernfelt, *Diagrammatology. An Investigation on the Borderlines of Phenomenology, Ontology, and Semiotics*, Springer, Dordrecht 2007.
P. Studtmann, *Aristotle's Categories*, in: The Stanford Encyclopedia of Philosophy, E. N. Zalta (ed.), Spring 2021.
A. Thomasson, *Categories*, in: The Stanford Encyclopedia of Philosophy, E. N. Zalta (ed.), Summer 2019.
F. J. Varela, E. Rosch, E. Thompson, *The Embodied Mind: Cognitive Science and Human Experience*, MIT Press, 1991.
J. A. Vieira, *Semiótica, sistemas e sinais*, PhD thesis. São Paulo: Pontifícia Universidade Católica, Pós-graduação em Comunicação e Semiótica, 1994.

ABOUT THE AUTHORS:

Eduardo Camargo — PhD, Post-Doc Researcher, DCA-FEEC-UNICAMP, Av. Albert Einstein, 400 13083-852, Campinas, SP, Brasil.
Email: cepca- margo@gmail.com

Ricardo Gudwin — PhD, Professor, DCA-FEEC-UNICAMP, Av. Albert Einstein, 400 13083-852, Campinas, SP, Brasil.
Email: gudwin@unicamp.br

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Mariusz Mazurek

# THE PROBLEM OF EXISTENCE OF VIRTUAL OBJECTS FROM THE PHILOSOPHICAL PERSPECTIVE

doi: 10.37240/FiN.2022.10.zs.6

***ABSTRACT***

I consider the problem of existence of virtual objects, mainly their mode of existence, while omitting the issue of the criteria of their existence. I present and analyze the concepts of modes (forms, kinds) of existence of virtual objects proposed in the literature of the subject, and then I demonstrate my own position on the issue. My position on the existence of virtual objects has certain points coinciding with the already postulated views, but at the same time it differs from them in some basic aspects. In my view virtual objects are "born" in human individual consciousness as the objects of specific creative states of the mind. So initially they are private objects belonging to the individual subjective sphere. However, their final and ready forms emerge in complex processes of objectifying and autonomizing the respective private conscious states and their objects. In these processes the private objects are transformed into objects intersubjectively accessible and existing in the collective cultural sphere. In both their forms, initial and final, virtual objects are non-material entities: first subjective, then objective. The ontic status of virtual objects is very similar to the status of intangible ideas and all the non-material objects created by the human mind. The main difference consists in that virtual objects are expressed by the use computers programmes, while other non-material objects created by human beings are expressed by use of words, pictures, literature and art works, etc.

**Keywords**: ontology, virtual existence, virtual objects, artefacts, fictions.

## VIRTUAL OBJECTS AS IMMATERIAL ENTITIES. MICHAEL HEIM'S POSITION

Michael Heim in his work *The Metaphysics of Virtual Reality*[1] describes and analyzes in detail the concepts related to virtual reality, such as: virtual, virtual reality, and virtual environment or virtual world. These terms are related to Heim's concepts of virtual entities and the primary world. When

[1] M. Heim, *Methaphysics of Virtual Reality*, Oxford University Press, Oxford 1993.

presenting Heim's concept, it is worth recalling the etymology of the very notion of virtuality, which comes from the Latin "*virtus*"—meaning virtue, strength, valor. "Virtual" means not existing in the empirical reality, but can exist—being potential and possible. The concept of virtuality in its general meaning is related to the auxiliary terms, such as: apparentness, imaginability, invisibility. The meaning of adjective "*virtualis*" would correspond to the meaning of today's adjective "potential," i.e. carrying the possibility of realizing some action (both adjectives contain Latin stems meaning power and strength—*virtus* and *potentia*).[2]

The essence of the virtual object is that it does not exist in a material way, but still functions in the reality accessible to the human senses. The virtual character of an entity or object means that its image or effects can be perceived by our senses, but not the entity or object as such. With the emergence and spread of computers the term "virtuality" began to be commonly used in the context of information technology rather than philosophy. This notion is particularly related to the development of the so-called virtual memory. In the case of computers, e.g., virtual memory can be part of the RAM. The expansion of memory does not require an additional space on the hard disk when using it. A virtual disk can be used in the same way as a hard disk, but it does not have its physical limitations. As computers have continued to evolve, especially the spread of the Internet, the term "virtuality" has expanded its meaning. Analogous to the concept of virtual disk, any entity or object is referred to as virtual when it functions in a manner devoid of the dimension of materiality.

The ontological consequences of this state of affairs directed the author of *The Metaphysics of Virtual Reality's* attention from computer science to philosophy, from which the term "virtuality" is derived.[3] A special role in Heim's analysis falls to issues central to the dispute over universals that took place in the Middle Ages. As Heim writes, the debate about virtual existence runs throughout the history of philosophy, but it gained particular significance in the writings of John Duns Scotus, whose views were a response to the system of Thomas of Aquinas, who referred to St. Augustine.

According to Heim, Duns Scotus created the discussed notion of virtuality (Latin: *virtualiter*), which is particularly important for understanding the theory of reality created by the medieval philosopher. It refers, in the writings of Duns Scotus, to the way in which form is connected with the physical attributes of things.[4] Thomas of Aquinas argued that the existence of extra-

---

[2] A. Pawłowski, *Wirtualizacja – historia i próba rekonstrukcji pojęcia* [Virtualization—History and an Attempt to Reconstruct the Concept], in: L. W. Zacher (ed.), Wirtualizacja problemy, wyzwania, skutki [Virtualization Problems, Challenges, Implications], Poltext, Warszawa 2013, p. 12.

[3] M. Heim, *Methaphysics of Virtual Reality*, op. cit., p. 132.

[4] Ibidem, p. 132. According to Heim, it was not until the Renaissance that the exclusive status of reality was first attributed to things perceivable by the senses. However, modern science rejects such conditioning of the definition of what is real, proving that the component and basis of reality are things and phenomena intangible to the senses, such as elementary particles or energy.

sensory entities (e.g. God) could be inferred from the facts of sense experience. On the other hand, Duns Scotus, who believed that philosophy should first of all deal with the generalized notion of being, was of the opinion that the knowledge of extrasensory being can be reached only without the testimony of senses and reason, through direct cognition of the essence of this being, which is infinite and brings into existence a universe of finite beings. Being, in Scotus' view, can refer both to God and to the universe of created beings.[5] Thus, in Duns Scotus' view, neither the testimony of the senses nor rational cognition can determine whether any object is called being or whether its existence is rejected. Duns Scotus never separated the notion of essence from existence, as Thomas did, and the notion of being in Scotus' view had a wide meaning—it included "everything that is not a thing," i.e. according to Heim's commentary, everything that exists in any way, also in a disembodied form, i.e. virtually.

According to Heim, the distinctive feature of the contemporary understanding of the term "virtuality" is specifically interdisciplinary. The concept was originated and functioned within philosophy, then entered computer science, and now it returns to philosophy. For this reason it has at least two basic meanings, which—however, in the language of different disciplines—seem to define the same phenomenon. Referring to Plato's thought, and especially inspired by Duns Scotus' definition of being, Heim combines philosophical and IT traditions of understanding the concept of virtuality and postulates calling objects existing in virtual worlds as virtual entities. Generally understood entities in his view are "all objects that can be registered as ontologically present or influencing the world."[6] Starting from this definition, Heim further narrows it down for the purposes of philosophical reflection on virtual reality. The entities that can be observed in this specific environment, which, as Heim emphasizes,[7] need not reflect any entities that exist in the material world, are all virtual objects (avatars, pictorial virtual representations, the so-called agent, i.e. an autonomous *software*-like object that is active in virtual worlds and can spontaneously change, evolve, or "learn").

On the other hand, virtual reality according to Heim, "is [...] a specific experience that gives the participant the impression of being in a different place than the one in which his body is currently located."[8] Heim associates the issue of virtual reality with computer technology, treating virtual reality as an area where processes similar to those occurring in empirical reality can take place through the use of three-dimensional digital graphics and electronic devices.[9]

---

[5] Ibidem, p. 117.
[6] Ibidem, p. 151.
[7] Ibidem, p. 147.
[8] M. Heim, *Metaphysics of Virtual Reality*, op. cit., p. 147.
[9] M. Heim, *Virtual Realism*, Oxford University Press, Oxford 2000, p. 6.

Heim indicates the following characteristics of virtual reality: (1) artificiality—conceived in terms of human presence in cyberspace, e.g., when the user's body and reactions are correlated with computer-generated images to give the impression of presence in the virtual world; (2) interactivity, i.e., the ability to engage into the virtual environment, e.g., by moving the cursor on a computer screen; (3) immersion—namely, using the computer to stimulate sensory experience, e.g., putting a helmet on the head with a screen on which a three-dimensional image appears. At the same time, the way in which the sense of presence and immersion simultaneously interpenetrate remains an open question of research on virtual reality; (4) communication in the network—consisting, among others, in the fact that different users can "enter" at the same time; (5) telepresence (telepresence; from Greek "tele"—"at a distance")—i.e., "presence at a distance"—an operation as a result of which the user feels present in a simulated area of virtuality, although he remains physically present in the material world, while the devices remotely transmit his actions.[10]

A certain shortcoming is that Heim's rather detailed characterization of virtual reality and the objects within it evades—as it seems—a definitive statement what modus of existence they have. It is not even known whether virtual objects are material (although strong suggestions indicate that they are not), and if the supposition of immateriality is correct, then the question arises as to where they are located, where they are present, and how they affect a material user (human). These and other problems that arise here cannot be solved by using the philosophical *instrumentarium* that Heim introduces. On the one hand, Heim invokes Plato's complex and rich metaphysics, but on the other hand, he does not effectively "adapt" it to the concept of virtual objects. The point is that the reference to Plato suggest that virtual objects are something like Platonic ideas. But it is only a suggestion, which is immediately objectionable: a cardinal difference is, among others, that virtual objects are created by man or by computer, while Platonic ideas are eternal, subjective, not founded anywhere.

## DO VIRTUAL OBJECTS EXIST IN REALITY?

As regarding its *mode* of existence, virtual reality is treated in opposition to material reality (the natural world), thus the problem of virtual reality's existence is related to one of the oldest ontological problems, i.e. the issue of distinguishing what is real and what is not real. Speaking of real existence, it is meant—as it seems—that what exists in external reality in relation to the mind of the subject creating or operating virtual objects, and thus—as a philosopher from the transcendental current would say—in reality transcendent to the subject.

[10] M. Heim, *Metaphysics of Virtual Reality*, op. cit., pp. 109–110.

According to the definition contained in the Polish Scientific Publishers (PWN) Encyclopedia, virtual reality is understood as:

> "... computer programs synthesizing sensations received by human senses (most often sound and image, but also touch), e.g. in flight simulators or computer games. In virtual reality systems, communication with the computer takes the form of visual (creating realistic, stereoscopic images of the simulated environment using computer graphics), audio, and tactile (using physical force to move in and control the simulated environment and to move simulated objects)."[11]

Here the question arises: Does the virtual reality understood in this way have to be created by computer? Is it reasonable to limit virtual objects only to computer creations? Is it advisable to follow a specific set of hardware and software, chosen arbitrarily, in order to define what virtual reality is? It seems that such definitions may become outdated with the development of computer technology and the emergence of new devices. Moreover, they omit the user, who not only handles the computer (operates it uncreatively), but above all creates virtual reality.

There is quite a large consensus of opinions that virtual reality is not a completely new phenomenon. New are only the means (keyboard, screen) by which we obtain, record and transmit information to others in this reality. Jeri Fink, author of *Cyberseduction: Reality in the Age of Psychotechnology*, rightly notes that: "... virtual reality is not a revolution but an evolution, a space occupied by humans since the first awareness of the qualitative difference between mind and body."[12] Theatrical performance, art, literature, and cinema are also ways of generating virtual reality (different from the "ordinary" reality experienced every day), but constructed with the use of other, traditional props. Fictional characters from books, movies, myths, and cultural symbols are also immaterial objects of reality, but constructed with the use of other, namely traditional tools (pen, pencil, paper), making more use of social and cultural contexts. The computer, sheet of paper, typewriter are only instruments, technical ways of expressing the thoughts of the entity that creates these objects, the same tool as the pen. The basis of all this is the human mind that invents and creates them.

If it is assumed that virtuality has little to do with the physical dimension of reality, it is perhaps much easier to relate it to its ideal or conceptual dimension. Fink expresses such suggestions when he writes:

> "... the virtual is generally defined as something that exists in the mind without reference to any physical fact, form, or feature. Virtual images are the

---

[11] https://encyklopedia.pwn.pl/haslo/wirtualna-rzeczywistosc;3996681.html (accessed: 05.12. 2021).

[12] J. Fink, *Cyberseduction: Reality in the Age of Psychotechnology*, Prometheus Books, New York 1999, p. 16.

> product of human creativity, insight and imagination. They emerge from conscious or unconscious processes, the activity of which constructs mental images."[13]

Fink also emphasizes that:

> "... a proper definition of virtual reality must go beyond simulations. It resembles looking at two mirrors reflecting each other—it is possible to see an endless series of similar images disappearing into infinite space. Logic suggests that this is a flat, hard piece of glass. It actually does not look flat. It does not feel like something flat. But if we touch it, we find a cold, hard surface with no depth, color or philosophy. What is more real? Glass or this image?"[14]

It may be assumed, following Fink, that virtuality is much more related to the ideal than to the material (physical). As long as we are dealing with a private, individual "mental image" of something that does not exist in any other way (already or yet), the "virtual" can mean the same thing as the "ideal," in the subjective sense. On the other hand, where we are dealing with already objectified creations of imagination, intersubjectively accessible and reproduced, for example via the Internet in any number of copies, which—what is important—can be contacted in an interactive way, virtual objects—as can be suspected—acquire additional properties or change their status. They become intersubjectively accessible, non-private objects, which is not the case with subjective mental images in the mind of their creator, i.e. in the immanent sphere.

## VIRTUAL OBJECTS AS MIRROR REFLECTIONS

The subject literature includes the view that virtual objects are in many ways similar to mirror reflections.[15] In such case, the criterion for distinguishing the virtual from the real should be found in considerations of distinguishing real objects from their mirror images. The issue of distinguishing real objects from their mirror images has been considered by philosophers for a very long time, definitely longer than the concept of virtual reality exists. Already Plotinus in the *Enneads*[16] analyzed the problem of mirror reflections and claimed that in most cases it is quite easy to distinguish a reflection from a real object. However, he agreed that under certain condi-

---

[13] Ibidem, p. 22.

[14] Ibidem.

[15] D. Stanovsky, *Virtual Reality*, in: L. Floridi (ed.), The Blackwell Guide to the Philosophy of Computing and Information, Black-well Publishing, Oxford-New York 2004, pp. 167–177.

[16] Plotyn, *Enneady* [Enneads], vols. I–II, A. Krokiewicz (trans.), PWN, Warszawa 1959, pp. 204–269.

tions (if the reflection in the mirror is permanent and the mirror alone cannot be seen) these reflections can deceive us.

It is easy to distinguish an object from its image in the mirror. This is because such a reflection is transient, fleeting, temporary, impermanent in time, and sometimes even inconsistent with other perceptual impressions. Moreover, in most cases the mirror does not go unnoticed; its structure, which reflects light, and its frame are visible. Moreover, the objects in the mirror are always depicted on a plane, and thus touching them makes it possible to distinguish the object from its reflection. It seems that virtual reality is more complex than mirror images and in some aspects different. The objects in it are not usually confined to a two-dimensional graphic, as is the case with the image reflected in the mirror, nor are they impermanent – they can last as long as real objects or events. Moreover, the mirror reflects only objects that exist in the present, while virtual reality objects may not have their real counterparts, they may exist only in virtual reality. Fictional characters in computer games do not copy any real objects, but are only creations of the human mind realized with computer techniques. Therefore, the concept, semi-metaphorical, of mirroring is not a satisfactory solution to the issue regarding the existence of virtual objects.

## VIRTUAL OBJECTS AS SIMULATIONS

The word "simulation" is derived from the Latin *simulatio*, which means "pretense," "a false representation of reality in order to mislead someone."[17] Colloquially, simulation is understood as imitating, mimicking, or replicating some original. In a particular case, simulation can be manipulative, when the imitated behaviors or roles are intended to falsify the image of reality. In a colloquial sense, simulation is merely an effort to fully conform to the original.

Virtual objects are often treated as computer simulations of real objects.[18] Simulation is then defined as an approximate reproduction of a phenomenon or behavior of a given object by means of its model. In this sense, a special (modern) kind of model is a model written in the form of a computer program, but it sometimes happens to use a physical (not virtual) model at scale.

It is reasonable to claim that computer program is a scheme of thinking, possibly of inference, of processing information realized with the use of

---

[17] Z. Rysiewicz, *Słownik wyrazów obcych* [Dictionary of Foreign Worlds], Warszawa 1955, column 694.

[18] A. Latawiec, *Rola symulacji w kreowaniu świata wirtualnego* [The Role of Simulation in Creating the Virtual World], in: A. Kiepas, M. Sułkowska, M. Wołek (eds.), Człowiek a światy wirtualne, wirtualne [The Human Being and Virtual Worlds], Wydawnictwo Uniwersytetu Śląskiego, Katowice 2009, pp. 50–58.

a computer. It could also be argued that it is not human thinking, but computer thinking—artificial intelligence. For example, a virtual object is a bridge design formed in a given program, but not, as it seems, that such an object is a representation of a program. A computer program is a tool (scheme for producing) virtual objects. This object is not a simulation of any existing object, but a design of such an object. In other words: it is a simulation of a potentially existing object.

Apart from the colloquial understanding of simulation in science, the term has many other meanings. Anna Latawiec divides it into two groups.[19] In addition to the already mentioned reproduction of the original material object, usually derived from empirical reality, simulation is referred to as a method of investigating reality by a specific algorithm. It consists in verifying or discovering features of reality using its immaterial model in virtual reality. In order to verify a certain phenomenon or process, usually not directly verifiable (e.g. the danger of spreading a virus), their immaterial model is created, which is subjected to verification in a computer environment (after writing an appropriate program, running it, verifying and interpreting its results). It is a kind of experiment in which the features of reality, but not of the virtual world, are discovered—this kind of simulation occurs in the imagination, e.g. at the creation of a research project.

In relation to simulation in the first sense (representation of reality), a man cannot act as a creator of virtual reality, but only as a reproducer. However, in relation to simulation in the second sense (creating a model and experimenting on it), the man is a creator of the virtual world—such a situation occurs in the case of film scripts, works of art in the project phase, musical works, computer games, thoughts, imaginations realized by means of visualization. The virtual world understood in such a way is a work of man, but the source of its creation is not only technology and knowledge about the known reality, but also the creator's own experience, the world of dreams, imagination and fantasies. Due to the fact that our knowledge about reality changes over time, the virtual reality evolves as well. From this we can conclude that the primary source of virtual reality is empirical reality (or the world of abstract objects, concepts, ideas) and the secondary source is the world of thought.[20]

Simulation is also defined—more specifically, more narrowly—as "the reproduction of properties of real objects in a digital environment,"[21] where "a perfect simulation is one that makes it impossible to know that we are

[19] Ibidem, p. 54.
[20] Ibidem, p. 56.
[21] J. Gurczyński, *Czym jest wirtualność. Matrix jako model rzeczywistości wirtualnej* [What is Virtuality. Matrix as a Model of Virtual Reality], Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, Lublin 2013, p. 125; See: D. Chalmers, *The Virtual and the Real*, Disputatio, 9 (46), 2017, pp. 309–352.

dealing with a simulation."[22] The concept of simulation is most often associated with systems called simulators (flight simulators, driving simulators), in which a faithful representation of reality is even essential.

If simulation is understood as imitating, mimicking or replicating some original, then it becomes problematic to define the case when there is no such original, i.e. when virtual objects or events do not have their real counterparts. As Richard Crandall and Marvin Levich[23] note, in virtual experiences we are dealing with real simulations and sometimes with fictional ones. A flight simulator—as its name implies—simulates actual flight. In contrast, experiences with fictional characters in virtual reality have little to do with experiencing reality (realness). It is also pointed out that what is attempted to be simulated in virtual reality are not only the objects as such, but also the occurrence between the perceived object and the cognitive entity of certain conditions under which the entity will experience it in a manner similar to that in which it experiences in reality.

Assuming that virtual objects are simulations of real objects, new questions arise that need to be answered. How and where do virtual objects arise? What is their ontic status? Are they as real as the objects of the world around us? From a materialist point of view, virtual reality is just a product of circuits and wires. However, it can be understood more broadly, not limited to the realm of technology: "... the virtual world consists not only of computer productions [...] but also of the world of art, movies, music, computer games, research projects, thoughts, and imaginations realized through visualization.[24]

It is worth noting that this understanding of the simulation concept is not a new phenomenon. Simulations of certain phenomena are simply well-known representations from empirical sciences. The way they are created—by means of a computer—is not particularly important.

Moreover, simulations understood as representations of phenomena that do not exist yet but will exist in the future (either as human creations or as predictions of future events), also in this computer simulation, are not a new phenomenon at all. Earlier, when there were no computers, weather forecasters used a piece of paper and a pencil, but the principle was the same: they took the current weather condition, wind directions, pressure levels and applying the laws of physics (more precisely geophysics, including meteorology) when calculating, they deductively gave the weather condition for the near future. Therefore, the computer plays here the role of only a faster and more efficient calculating and thus forecasting instrument.

---

[22] Ibidem, p. 126.

[23] R. Crandall, M. Levich, *Virtual Reality, And All That*, in: A Network Orange. Logic and Responsibility in the Computer Age, R. Crandall, M Levich (eds.), Springer, New York, 1998, pp. 85–107.

[24] A. Latawiec, *Rola symulacji w kreowaniu świata wirtualnego* [The Role of Simulation in Creating the Virtual World], op. cit., p. 53.

Finally, it is worth noting that in recent years quite popular is the concept of augmented reality[25], referring to computer mediated experience in which the signals coming from the real physical world are supplemented by virtual signals in real time, performing simulation functions. Appropriate synchronization and coordination enable, e.g., a surgeon to operate on a live patient using a computer simulation of invisible internal organs. The opposite situation is also possible: a medical student, using an authentic scalpel, can operate on a virtual patient and thus gain experience.

## VIRTUAL OBJECTS AS FICTIONS

Virtual objects (particularly characters from computer games, books, and myths) share many characteristics with fictional objects.[26] Fictional objects are usually understood as characters, objects or events that appear in myths, literary or film works (Zeus, Pegasus, Sherlock Holmes, the flying carpet).[27] Fictional objects are contrasted with real, actual objects (e.g., historical figures).

As Jacek Gurczyński notes:

> "... fictional objects are necessarily non-existent (real) objects, but this claim should be distinguished from the assertion that fictional objects have no ontological status. The basic assumption that allows to grant fictional objects some existential status [...] is Franz Brentano's thesis that every act of consciousness has an object, i.e., in other words, every act of consciousness is intentional. Whenever we speak, we talk about something, our fantasies are about something, we look at something, we think about something—any conscious experience is always 'consciousness of something'."[28]

The property that distinguishes virtual objects from other fictional objects is the interactive mode of telepresence of the former. Interactivity is a broad phenomenon, occurring not only in relation to computer technologies. We also deal with it during videoconferences, in radio and television programmes using, e.g. telephone or audiovisual communication with the audience.

---

[25] See: T. Metzinger, *Why Is Virtual Reality Interesting for Philosophers?* Frontiers in Robotics and AI, 5, 2018, pp. 101-120.

[26] https://plato.stanford.edu/entries/fictional-entities/ (accessed: 05.12.2021).

[27] https://plato.stanford.edu/entries/nonexistent-objects/ (accessed: 05.12.2021). It is worth to note here that the authors do not refer at all to older, classical concepts of existence—which is surprising—and they conduct their considerations mainly on the basis of works from the last few decades, which are limited to modern concepts, very specialized, above all, deeply entangled in logic, and not in classical philosophical solutions. Meanwhile, the problem of existence has been present in philosophy since the time of Plato.

[28] J. Gurczyński, *Czym jest wirtualność. Matrix jako model rzeczywistości wirtualnej* [What is Virtuality. Matrix as a Model of Virtual Reality], op. cit., p. 183.

Piotr Sitarski describes it in a similar way, claiming that the most important distinctive feature is the interactive way of obtaining the feeling of telepresence, thus emphasizing that such a feeling may be obtained in many ways. This author writes:

> "Thus, the specificity of virtual reality is that the impression of being in another world is achieved through the interactivity of the system. The user feels present in the mediatized environment not as a result of sensory illusion or narrativization, but as a result of the possibility of action, analogous to action in the normal world. Virtual reality is based on interaction that immerses the participant in a fictional world."[29]

Virtual objects, like fictional and real objects, are intersubjectively accessible—they are not private experiences of the entity, such as dreams. Perceived from the outside, they are also characterized by two levels of determination. Consider, for example, computer game characters. They are characterized by two types of features: internal as elements of the virtual world (game world)—appearance, character traits, skills, gender, etc.—and external, attributed to them from the perspective of reality (outside the game)—being an object of the game, being virtual, being created by the creators of the game. According to Gurczyński, this two-level determination can be generalized: "whenever the world s2 is superstructured over (is ontically dependent on) the world s1, then objects from the world s2, perceived from the level of the world s1, will be characterized by a two-level emplacement."[30] Then this characteristic is a common property of fictional and real objects.

The objects of computer games change as they are manipulated by the player (according to the rules set in the program that defines the game rules). A computer game can be compared to a movie (moving pictures). One of the main game attractions is the ability to stop time at any time and speed it up to achieve certain goals faster. Time can flow just like the real one, but it can also be reversed, allowing you to cancel selected events that have already occurred.

However, there is a difference—games, unlike movies, are to a greater extent interactive, i.e. their user constantly participates in their course and has influence on what is happening. The user shapes the game course within the set rules (he can make only such interventions in the action that are designed). The person watching a movie is passive, while the player continually creates the plot of the computer game, because games, unlike movies, are

---

[29] P. Sitarski, *Rozmowa z cyfrowym cieniem. Model komunikacyjny rzeczywistości wirtualnej* [Conversation with Digital Shadow. A Communication Model of Virtual Reality], Rabid, Kraków, 2002, p. 42.

[30] J. Gurczyński, *Czym jest wirtualność. Matrix jako model rzeczywistości wirtualnej* [What Is Virtuality. Matrix as a Model of Virtual Reality], op. cit., p. 199.

interactive, their very essence is the participation of players (their participation in virtual reality), and this causes players to accept the game and virtual reality in general as equally real and of the same ontic type as their everyday lives. It is worth noting here that we come to a certain paradox, as from the point of view of the criterion of interactivity—that what is virtual is sometimes more real than that what is physical.

In the case of movies, the carrier of content (storyline) is film stock, and nowadays it is a camera and an electronic medium (e.g. a computer disk or a telephone disk). The material medium of a computer game, on the other hand, is a computer or other device (telephone). They are the material means of transmission and dissemination of content, film plot, etc.

## VIRTUAL OBJECTS AS ARTIFACTS

The identification of the category of virtual objects with the category of artefacts usually means that virtual objects as artefacts belong to the artificial world created by man who is separate from the natural world. Therefore, it should be assumed that they are also separate from nature.[31] This position is proposed, among others, by Józef Lubacz and Krzysztof Brzeziński, for whom "virtuality and its objects are artifacts, i.e. intentional creations of people."[32] Despite the fact that the ways of creating virtual objects and material artifacts are similar, i.e. man-made, the ontic status of both types of objects seems to be different.

Risto Hipelin states that the artifact is "an object that has been created or produced for a given purpose."[33] Etymologically, the word "artifact" comes from Latin words "*arte*" ("skill," also "art," "technique") and "*factum*" ("to make"). Artifact is a concept that is increasingly being used in both aesthetics and technology, and is more likely to refer to material objects (e.g., works of art, telephones). Recently, the concept of artifact has been introduced into epistemology;[34] in general, it becomes more fashionable, and its scope successively increases. The *Mirriam-Webster Dictionary* states that an artifact is a characteristic product of human activity. Artifacts may include tools,

---

[31] M. Krueger, *Artificial Reality II*, Addison-Wesley Publishing Company, Massachusetts 1991; E. Margolis, S. Laurence, *Creations of the Mind. Theories of Artifacts and Their Representation*, Oxford University Press, Oxford-New York 2007; P. Sitarski, *Rozmowa z cyfrowym cieniem. Model komunikacyjny rzeczywistości wirtualnej* [Conversation with Digital Shadow. A Communication Model of Virtual Reality], op. cit., p. 42; M. Heim, *Virtual Realism*, op. cit., p. 6; idem., *The Metaphysics of Virtual Reality*, op. cit., p. 131.

[32] K. Brzeziński, J. Lubacz, *Skąd się biorą przedmioty wirtualne* [Virtual Objects and Where They Come From], in: Przedmioty wirtualne [Virtual Objects], P. Stacewicz, B. Skowron (eds.), Virtual Objects, Warsaw University of Technology Publishing House, Warsaw 2019, pp. 11–23.

[33] https://plato.stanford.edu/archives/sum2018/entries/artifact/ (accessed: 05.12.2021).

[34] M. Trybulec, *W stronę epistemologii artefaktów poznawczych* [Towards the Epistemology of Cognitive Artifacts], Filozofia i Nauka. Studia filozoficzne i interdyscyplinarne, 9, 2021, pp. 195–223.

works of art, especially archaeological objects.[35] Artifacts are also defined as something created by humanity, which does not belong to nature, which is not nature.[36]

If we accept the standard definition of artifacts as material objects, then identification of virtual objects as artifacts seems to be incorrect (artifacts are material, and virtual objects are immaterial). In the broadest sense of the term "artifact," the class of artifacts is rich and diverse. As a result, including virtual objects in this class does not tell us much, especially if it is not clear whether the artifacts are material or immaterial as well. If man-made immaterial objects are also included in the set of artifacts, then it can be argued that virtual objects are artifacts, and form its special class. However, the mere inclusion of virtual objects in the (extended) set of artifacts is not very informative, because—it does not indicate the specificity of virtual objects that differentiates them from other artifacts, in other words, it does not specify its subtype among the whole kind of artifacts. It would be necessary here to specify the attributes that differentiate them from other artifacts. Otherwise, saying that virtual objects are artifacts (in these extended senses) is to state the obvious, which in fact provides little information about the existence of virtual objects. If, on the other hand, only specific, i.e. man-made, material objects are considered artifacts, then virtual objects are not artifacts at all. Such a classification is erroneous.

Already Aristotle divided and described the difference between what exists naturally and artificial creations created by other causes, noting that:

> "'By nature' exist animals and their parts, as well as plants and simple bodies, such as earth, fire, air, and water—as these and the like are said to exist 'by nature.' It further appears that all the things mentioned are different from those which are not products of nature. For every object of this kind bears a principle of motion and rest: some due to place, others due to growth and decay, and finally others due to qualitative change. On the contrary, a bed, a robe, and other things of this kind, insofar as they are entitled to such general predicates, and insofar as they are products of craftsmanship, do not reveal any natural tendency to change. However, if by chance they are made of stone or of earth or from a combination of both, then they reveal such a tendency, but only in this respect: indeed, 'nature' is the principle and intrinsic cause of movement and rest in things in which it exists intrinsically and not accidentally."[37]

Although the contemporary concept of artifact can be understood more broadly than Aristotle proposed (who, of course, did not use this term), his

---

[35] https://www.dictionary.com/browse/artefact (accessed: 05.12.2021).

[36] https://www.merriam-webster.com/dictionary/artifact (accessed: 05.12.2021).

[37] Arystoteles, *Fizyka* [Physics], Księga II [Book II], K. Leśniak (trans.), Warszawa 2010, pp. 87–88.

observation, pointing to two kinds of material things, has not essentially lost its validity.

Currently, attention is paid mainly to the problem of intention and subject (author) than to the way of existence, although both approaches are related. Risto Hilpinen, referring to the term proposed by Wendell Oswalt, naturefacts, which refers to "things existing by nature," proposes a dichotomy similar to the Aristotelian one, pointing to the existence of artifacts that are the result of human activity. He notes that the creation of something requires an intentional action—with the term "intentional" being understood as a conscious action with a specific purpose.

> "Technical artifacts, such as typewriters, hammers, copying machines, or computers, differ from social artifacts—such as laws or money—in that the realization of their function depends fundamentally on physical structure. They also differ from physical or natural objects in that they are produced intentionally and used by human beings to achieve certain goals."[38]

Artifacts so defined have two essential characteristics: (1) they are material objects the properties of which determine the performance of a function, and (2) they have been produced for certain purposes. Therefore, it can be stated that they have a dual nature. It seems to be a common belief in the literature that artifacts have such a dual nature.[39] On the one hand, they are material, their structure providing the possibility to perform certain functions, while on the other, these functions refer to something immaterial—human intentions. It should be noted that this identification of artifacts as both material and intentional objects is problematic. Something that is created intentionally can be both material and immaterial. It is possible to narrow the extensiveness of an artifact to material objects only, as proposed, e.g., by Wendell Oswalt in one of his definitions, according to which an artifact is "an end product resulting from the modification of a physical mass in such a way that it can fulfill a purpose and become useful." Such a general definition that an artifact is "an object created intentionally or for a specific purpose" raises problems. An object that has been created intentionally can be both material and immaterial in nature. From the anthropological research point of view, which was the subject of his interest, it may be adequate, but from a philosophical perspective it seems far too narrow. It is possible to make a general statement that what cannot be defined as a natural kind (understood broadly) is an artefact, thus also an immaterial product.

---

[38] P. Kroes, A. Meijers, *The Dual Nature of Technical Artefacts*, Studies in History and Philosophy of Science, 37 (1), 2006, p. 1.

[39] L. R. Baker, *On the Twofold Nature of Artefact*, Studies in History and Philosophy of Science, 37, 2006, pp. 132–136; W. Houkes, A. Meijers, *The Ontology of Artefacts: The Hard Problem*, Studies in History and Philosophy of Science, 37(1), 2006, pp. 118–131; P. Kroes, A. Meijers, *The Dual Nature of Technical Artefacts*, op. cit., pp. 1–4.

Amie Thomasson highlights this point, indicating that the role of human intention in the process of creating artifacts is different from that of creating social and institutional objects. Unlike the latter, the existence of an artifact does not assume any collective intention[40], but it is necessary that the action (leading to the creation of artifact) be conscious. Although Thomasson does not deny the possibility of the existence of artefact's essence, he points out that in the case of artifactual kinds it is constituted precisely by the intention of their creators, which clearly distinguishes them from natural kinds.[41]

It is also worth noting that in contemporary analytical philosophy, the above problem is also discussed, as by A. Thomasson, in the form of referring to the terms "natural kind" and "social kind." The first of these terms is used in many contexts and its use is controversial. Basically, it refers to classes of naturally existing elements of reality, studied by natural sciences.[42] Artifacts can thus be considered social types in the sense that they are produced by society. It is easy to recognize that such a term is quite close to Aristotle's conception.

However, no universally accepted theory of artifacts has been developed that would combine both of their inherent aspects: material and intentional (functional). The greatest difficulty of existing concepts seems to be definition of the concept of artifact function. After all, on the one hand, if one understands function as possible in material objects, the question arises how this function is related to states of human mind. However, if to assume that function is a certain state of mind, then it exists only in the imagination of designers and users of artifacts. Then it is difficult to explain how it is related to the material substance that makes up a particular artifact. Hence, the notion of an artifact's function is a key concept for its description, linking the material and intentional domains.[43]

In summary, if it is assumed that artifacts are immaterial man-made objects (like all cultural creations), then the specificity of these objects is not given and the concept of artifice is not a differentiating feature. It is also false to assume that they have an author—they are often nameless creations that have been created by someone—it does not matter that they are human creations (for example, mythical figures do not have an author—they grow out of tradition and culture, were passed down through word of mouth, and then passed on by others—they do not have a specific author, and there are often many creators of an artifact).

In general, it is unacceptable to attribute a material form to virtual objects. It seems that this misidentification of virtual objects as artifacts results

---

[40] A. Thomasson, *Artifacts and Human Concepts*, in: Creations of the Mind, E. Margolis, S. Laurence (eds.), Oxford 2007, p. 52.

[41] Ibidem, p. 53.

[42] https://plato.stanford.edu/archives/win2016/entries/natural-kinds/ (accessed: 05.12.2021).

[43] P. Kroes, *Technical Artefacts: Creations of Mind and Matter. A Philosophy of Engineering Design*, Springer, Dordrecht 2012.

from confusing them with images (e.g., on computer screens), depictions in the material world.

> "The [virtual – MM] object is one of the elements of produced implementation. It can be passed on for further use in a permanent tangible form (an image in a frame), or in a form hidden in the construction of other elements of the implementation, and physically appear when these elements are used (an image on a computer screen, which first had to be turned on and run the appropriate program on it)."[44]

The quotation is a statement in favor of material existence of virtual objects. However, it is worth noting that the sense of this statement is not unambiguous. It can be interpreted as a statement in favor of permanent material existence of virtual objects (it has a permanent tangible form, the authors write) or conditional material existence (it becomes physical when it appears on the computer screen at the moment when we turn on the computer and run an appropriate program). Whereas when, it may be added, the computer is switched off, the virtual object loses its material modus of existence, but it is not known whether it passes into another way of existence. As can be seen, the declaration concerning material existence of virtual objects leads to difficulties, which—not difficult to show—are growing in the course of their further consideration.

## VIRTUAL OBJECTS AUTONOMIZED (OBJECTIFIED) PRODUCTS OF CONSCIOUSNESS

The standpoint on the existence of virtual objects that I propose has some convergence points with the views already presented in the subject literature, but I add important aspects to them. Signally, I consider the genesis of virtual objects in human individual consciousness, and their final, ready-made form emerging in the process of objectification of relevant consciousness states. This objectification is accompanied by the autonomization of virtual objects—they cease to be the private property of the individual human subject, who is the creator of the virtual object.

The ontic status of virtual objects is similar to the status of immaterial ideas that appear in the mind of the subject who produces them. It can be stated that an idea, which is an element of knowledge or its immaterial subject, represents an object that does not exist at the moment of its creation, but is an object, an entity, a potential or pure form in the sense already formed by the Greek philosophers.

The virtual object is created by the subject and in the first phase of creation process it is a subjective invention of the subject, a private object of his

---

[44] K. Brzeziński, J. Lubacz, *Skąd się biorą przedmioty wirtualne* [Virtual Objects and Where They Come From], op. cit., pp. 19–20.

consciousness, inaccessible to anyone except the creator, i.e. existing only in the private immanent sphere—as transcendental philosophers would say. Then there is an objectification process of this invention and, at the same time, its autonomization. This phase is based on the communication of the subjective creation to other subjects. In this communication, the desubjectivization of the virtual object takes place. In the course of intersubjective communication it penetrates into the consciousness of other subjects and thus loses its character of an exclusively private object, belonging only to the consciousness of its creator. It becomes the property of other subjects, participants in the communication process, penetrating into their consciousness, and finally a common, collective property. After dissemination and desubjectivation (i.e. after the information about the virtual object has been communicated to other subjects) the object of individual consciousness of the subject-creator becomes independent of the creator in a sense that its consciousness is irrelevant to the way it functions and exists. The virtual object exists and functions in an objective cultural reality. It becomes, in a way, a common property of many subjects, eventually also a public property, an object of collective consciousness or, in other terms, an object of the third world—if we use here the notion from Karl R. Popper's conception of three worlds.

When the creator of a virtual object makes the content of the virtual object accessible to other users through communication and dissemination, the object becomes objectified. Inspired by Popper's concept of three worlds cited above, it can be assumed that virtual objects, which in the first phase of their creation are private, subjective objects, autonomize and at the same time objectivize when they become available in the processes of communication by means of computers, including the dissemination of computer programs and content transmitted via the Internet and other computer information carriers. Therefore, in conclusion, there are simultaneously occurring and interrelated processes: desubjectivation, objectivization and autonomization. In these processes, the virtual object is finally formed from the private ideas of individual subjects.

Virtual objects are located in an intangible, objective, man-made world. It can be called the world of cultural objects (in the broad sense of the term "culture"). They affect human consciousness, and thus—by initiating human activity in the material sphere—the physical world. For example, virtual house designs are the basis for building real houses. Virtual objects in computer games influence the users of these games by, among others, increasing their imagination, teaching them, and inducing aggressive behavior.

Virtual objects are realizations of ideas appearing in the individual consciousness of their creators. At first they are immaterial objects, sometimes they are materialized (e.g. a house built according to its virtual design), often they remain immaterial (e.g. characters from computer games). Between

the idea and the immaterial object of this idea there is a relation of representation: immaterial (virtual) object represents the idea. However, when the object of this idea is materialized, it can be stated that this material object represents the idea and at the same time represents its virtual project. Taking this together, it can be seen that there are several mutually different relations of representation: between the idea and the (intentional) virtual object corresponding to this idea, between the virtual object and its material realization, between such realization and the idea.

In this case, representation is a relation between the idea and the corresponding object or, most often, the whole set of objects—both existing and not yet created, but only possible, in the sense that they are designed, i.e. have an idea or model. The set of objects representing a given idea (represented) should include both created objects and potential objects. Therefore the set of objects representing an idea is ontologically complex. It includes objects of two modes of existence: the objects existing at present in virtual reality and the objects existing potentially, which are to be created in the future. This set changes in time, i.e. it is temporal, unstable, changeable, expanding. In other words, objects representing a given idea constitute an open set. It includes realized objects, produced in immaterial form equivalents of the idea, and possible objects, not yet created. The complication is that the represented objects, or ideas, exist in the immaterial world, while the objects representing them exist in material reality and in potential reality.

Certain aspects of the view presented above, i.e. those which proclaim that the virtual world is the work of a man (and not of the computer) as well as the position that it is the human subject who creates virtual objects, can be found in the literature in Elisabeth Reid's views:

> "Virtual worlds do not exist in the technology used to represent them or solely in the mind of the user, but in the relationship between internal mental constructs and technically produced representations of those constructs. The illusion of reality lies not in the apparatus, but in the users' desire to treat the products of their imagination as if they were real."[45]

This is also pointed out by Tadeusz Miczka, who states: "virtual reality is, after all, an artificial creation that exists intentionally because a person wants to endow it with existence,"[46] as well as Anna Latawiec, who claims:

> "By virtual world I mean the image of reality created or discovered by man on the way of intellectual or technical simulation. The world understood in this way is a creation of man. It has its source in widely understood reality. By re-

---

[45] E. Reid, *Cultural Formations in Text—Based Virtual Realities*, PhD Thesis, University of Melbourne, Melbourne 1994, pp. 6–7.

[46] T. Miczka, *Czysta iluzja i testowanie realności: dwie rzeczywistości wirtualne – dwa uczestnictwa* [Pure Illusion and Testing Reality: Two Virtual Realities—Two Participations], in: Człowiek a światy wirtualne [The Human Being and Virtual Worlds Worlds], A. Kiepas, M. Sułkowska, M. Wołek (eds.), Wydawnictwo Uniwersytetu Śląskiego, Katowice 2009, p. 19.

> ality I mean both the empirical reality, which is accessible only to a limited extent to the observer, and 'reality itself,' i.e. existing beyond the observer's reach."[47]

However, in the aforementioned positions of Elizabeth Reid and Anna Latawiec, there is no explanation of how objects become public and in what reality they are located. I believe that some modification of Popper's conception of the third world and the hypothesis of how the process of transition from a private creation to an objectively existing object occurs is a necessary part of the image regarding the existence of virtual objects.

In conclusion, virtual objects are immaterial objects, which in the phase of their creation are private objects of individual consciousness of their creator, and finally—after their desubjectivation, objectivization and thus autonomization (from their creator) are objectively existing objects in immaterial cultural reality. Sometimes the way of existence of virtual objects is confused with the criterion of their existence. At the same time, some philosophers claim that there is no possibility to inquire what existence is (i.e., in fact, what is its modus), as it is an ontological mystery. This is, as it seems, an escape from the problem, and there are, after all, useful inspirations for it in the philosophy of various schools, since antiquity. It is possible to discuss only the criteria of existence, namely, the methodological or epistemic principles by means of which we affirm that something exists—thus not getting into the question of what this existence is and what it consists of. Ancient, medieval, and modern philosophy have developed various deep and elaborate classifications of existence modes.[48] This problem cannot be considered within the framework of a materialist ontology, which postulates the reduction of everything that exists to material objects.[49]

## REFERENCES

Arystoteles, *Fizyka* [Physics], Księga II [Book II], K. Leśniak (trans.), Warszawa 2010, pp. 87–88.

L. R. Baker, *On the Twofold Nature of Artefact*, Studies in History and Philosophy of Science, 37, 2006, pp. 132–136.

K. Brzeziński, J. Lubacz, *Skąd się biorą przedmioty wirtualne* [Virtual Objects and Where They Come From], in: Przedmioty wirtualne [Virtual Objects], P. Stacewicz, B. Skowron (eds.), Warsaw University of Technology Publishing House, Warsaw 2019, pp. 11–23.

D. Chalmers, *The Virtual and the Real*, Disputatio, 9 (46), 2017, pp. 309–352.

---

[47] A. Latawiec, *Rola symulacji w kreowaniu świata wirtualnego* [The Role of Simulation in Creating the Virtual World], op. cit., p. 52.

[48] A. Chmielecki, *Sposoby istnienia* [Modes of Existence], Filo–Sofija, 1 (2), 2002, pp. 7–21; W. Krajewski, *O podstawowym – i niepodstawowych sposobach istnienia* [On the Basic and Non-Basic Modes of Existence], Filozofia Nauki, 10 (1), 2002, pp. 67–82.

[49] The reductionist current dominates American philosophy of mind, which sets the tone for research on the problem of mind on a global scale. A similar trend can be observed in the philosophy of computing, in the research on the ontic status of virtual objects: they are either treated as (non-existent) fictions or attempted to be reduced to material objects—both these trends seem inappropriate.

A. Chmielecki, *Sposoby istnienia* [Modes of Existence], Filo–Sofija, 1 (2), 2002, pp. 7–21.
R. Crandall, M. Levich, *Virtual Reality, And All That*, in: R. Crandall, M Levich (eds.), A Network Orange. Logic and Responsibility in the Computer Age, Springer, New York, 1998, pp. 85–107.
J. Fink, Cyberseduction: Reality in the Age of Psychotechnology, Prometheus Books, New York 1999.
M. Heim, *The Metaphysics of Virtual Reality*, Oxford University Press, New York 1993.
____ , *Virtual Realism*, Oxford University Press, Oxford 2000.
W. Houkes, A. Meijers, *The Ontology of Artefacts: The Hard Problem*, Studies in History and Philosophy of Science, 37(1), 2006, pp. 118–131.
J. Gurczyński, *Czym jest wirtualność. Matrix jako model rzeczywistości wirtualnej* [What Is Virtuality. Matrix as a Model of Virtual Reality], Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, Lublin 2013.
W. Krajewski, *O podstawowym – i niepodstawowych sposobach istnienia* [On the Basic and Non-Basic Modes of Existence], Filozofia Nauki, 10 (1), 2002, pp. 67–82.
P. Kroes, A. Meijers, *The Dual Nature of Technical Artefacts*, Studies in History and Philosophy of Science, 37 (1), 2006, pp. 1–158.
P. Kroes, T*echnical Artefacts: Creations of Mind and Matter. A Philosophy of Engineering Design*, Springer, Dordrecht 2012.
M. Krueger, *Artificial Reality II*, Addison-Wesley Publishing Company, Massachusetts 1991.
A. Latawiec, *Rola symulacji w kreowaniu świata wirtualnego* [The Role of Simulation in Creating a Virtual World], in: A. Kiepas, M. Sułkowska, M. Wołek (eds.), Człowiek a światy wirtualne [The Human Being and Virtual Worlds], Wydawnictwo Uniwersytetu Śląskiego, Katowice 2009, pp. 50–58.
E. Margolis, S. Laurence, *Creations of the Mind. Theories of Artifacts and Their Representation*, Oxford University Press, Oxford–New York 2007.
T. Metzinger, *Why Is Virtual Reality Interesting for Philosophers?* Frontiers in Robotics and AI, 5, 2018, pp. 101–120.
T. Miczka, *Czysta iluzja i testowanie realności: dwie rzeczywistości wirtualne – dwa uczestnictwa* [Pure Illusion and Testing Reality: Two Virtual Realities—Two Participations], in: A. Kiepas, M. Sułkowska, M. Wołek (eds.), Człowiek a światy wirtualne [The Human Being and Virtual Worlds], Wydawnictwo Uniwersytetu Śląskiego, Katowice 2009, pp. 11–29.
A. Pawłowski, *Wirtualizacja – historia i próba rekonstrukcji pojęcia* [Virtualization—History and an Attempt to Reconstruct the Concept], in: L. W. Zacher (ed.), Wirtualizacja problemy, wyzwania, skutki [Virtualization Problems, Challenges, Implications], Poltext, Warszawa 2013.
Plotyn, *Enneady* [Enneads], vol. I–II, A. Krokiewicz (trans.), PWN, Warszawa 1959, pp. 204–269.
E. Reid, *Cultural Formations in Text—Based Virtual Realities*, PhD Thesis, University of Melbourne, Melbourne 1994.
Z. Rysiewicz, *Słownik wyrazów obcych* [Dictionary of Foreign Worlds], Warszawa 1955, column 694.
P. Sitarski, *Rozmowa z cyfrowym cieniem. Model komunikacyjny rzeczywistości wirtualnej* [Conversation with Digital Shadow. Communication Model of Virtual Reality], Rabid, Kraków 2002.
D. Stanovsky, *Virtual Reality*, in: L. Floridi (ed.), The Blackwell Guide to the Philosophy of Computing and Information, Black-well Publishing, Oxford–New York 2004, pp. 167–177.
A. Thomasson, *Artifacts and Human Concepts*, in: E. Margolis, S. Laurence (eds.), Creations of the Mind, Oxford 2007.
M. Trybulec, *W stronę epistemologii artefaktów poznawczych* [Towards the Epistemology of Cognitive Artifacts], Filozofia i Nauka. Studia filozoficzne i interdyscyplinarne, 9, 2021, pp. 195–223.

ABOUT THE AUTHOR — PhD, Institute of Philosophy and Sociology of the Polish Academy of Science, Nowy Świat Street 72, 00-330 Warsaw, Poland.

Email: mmazurek@ifispan.edu.pl

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Bogdan Popoveniuc

# PERSONAL AND MORAL IDENTITY IN THE 4th SPACE

doi: 10.37240/FiN.2022.10.zs.7

***ABSTRACT***

The 4th Space concept is a very challenging and puzzling one. The tremendous technological progress of Information and Communication Technologies (ICTs) or Computer-Mediated Communication (CMC), ubiquitous computing, and Extended Reality (XR) make the Gibsonian Cyberspace Matrix an imminent reality in the future. Although, some features can be made more salient, the structure, but most importantly, the effects of living in such environment for human consciousness and morality is almost impossible to predict. Hence, the requisite of a proactionary and comprehensive scientific and technical paradigm for designing the 4th Space, in order to facilitate the adaptation of human species to the brave new technological world, while preserving the humanness and humanism of the humans.

**Keywords**: 4th Space, cyberspace, Information and Communication Technologies (ICTs), rhizome, autopoietic systems, ubiquitous computing, ISelf, infraethics, dispersion of responsibility.

## INTRODUCTION

This article outlines an epistemological and conceptual framework for understanding the envisaged 4th Space as the "onlife" living environment. The 4th Space will be analyzed based on the homology and continuity between physical, psychological, social and cultural spaces. In the beginning, is highlighted the foundational role of the space in the genesis of the Self and personal identity. After this, it will be presented the hybrid nature of the 4th Space based on the latest conceptualizations draw from ICTs, technological extended reality (XR), virtualization, ubiquitous computing, and psychology. It is argue that the paradigm of autopoietic system is the most suitable for understanding the formation personal and moral identity within the 4th Space. In addition, the particular processes affecting the personal identi-

ty within the 4th Space onlife environment are discussed in terms of the rhizomatic model of development and using phenomenological approach. In the proposed framework, are made salient the limitations of the current scientific paradigm in terms of its inability to ensure control and safeguard the humanness from the blind consequences of technological progress. In the end, will be discussed some epistemological, methodological and educational challenges and are proposed paths for future researches.

The notion of 4th Space is definitely a romantic one, in a double sense. Firstly, because it is an idea that "release[s] from, or transcendence[s] over, ignominious or uncomfortable circumstances" (Smith, 1996, p. 6). It elevates our imagination above our mundane condition of beings who adapt themselves to an external environment. It feeds the need of self-improvement and self-development of human soul, its poietic sense of the world, because the world ceased to be a given thing in itself, but it is how we experience and co-create it. It is a promise of enhancing our social world and fulfilling our innate need for relatedness. Secondly, because its conceptualization doesn't withstand yet the rigors of social sciences analysis. Its conceptualization is still about a prospective reality. And this, is not just a limitation, but both a sign and a warning. A sign that, in our increasingly technological environment dominated by scientific paradigm of understanding the world, we need the humanities to keep our humanness alive. A warning, but also an opportunity, for the modern science paradigm to become aware of the necessity to acquire the near future and the forthcoming states of the world as a legitimate object of study. Otherwise, it risks being each time one step behind the social and technological advancement and its knowledge to be always outdated. Most important, it risks to be unappropriated and even harmful to our self-understanding.

## SPACE AS THE FOUNDATION FOR SUBJECTIVITY

The physical space is the condition *sine qua non* for living in the world that precedes thinking: hence it is very opaque to analysis. The cognitive development and the psychological space are build based on concrete operations and movements in physical reality (Piaget, 1952). Even though psychological space (and time) are subjective conditions of knowledge and mental life, they must have "objective" value, i.e. they are valid for all individuals, because they were shaped in the environment-organism interaction that necessarily imprinted on us the features of objective reality. It would be impossible for the human species to survive if its subjective forms constantly and completely distorted the data of the objective world" (Joja, 1971, p. 190).

Space perceived as given reality for a consciousness is not an empirical concept or representation because it is already presupposed at the ground of

things represented as being outside. It is a condition of possibility of all experience (*appearances*) given as outside reality and we cannot imagine there is no space (Kant, 1781). Although space, as a subjective condition of conscious sensibility, is a necessary representation at the ground of all experiences, its a *priori*-ness does not exclude a genesis and an unceasingly and unconsciously process of reinstating. The individual perceives and recognizes the world as a given environment and "materializes" it as stable or as given reality by manipulating objects and making movements in space. The sense of reality is given by the unconscious and continuous connection with the environment/objects.

At its origin, space is related with the human body. As John Eliot (1987) shows, the space as psychological reality is the result of a combination of visual perception, touch, and movements: sensory integration, perceptual awareness, bodily awareness and the coordination of physical movement, skillful performance of spatial tasks, conceptual experience that transcends the perceptual moment and provides direction to our behavior. The 4th Space concept deepens the psychological integration and upholds the defining role of psychological space for human life to the detriment of physical space. The 4th Space is a mental-virtual-physical space and from here the vital importance of and effect on psychological processes. The plasticity and dynamics of virtual space, that can be construct and change in unimaginable ways the context of communication and perception pose a great challenge for human psychic. Reality and its spatial aspect/dimension are our perceptual awareness of relationships surrounding us and our thinking about them are independent but related aspects of human cognition. Also, perception and cognition are largely independent, but still affect each other in systematic ways (Montemayor, Haladjian, 2017).

In this context, the question of the 4th Space announced itself from the beginning to be very intriguing. As it emerges both as environment and an extension of the psychological space (thoughts, memory, representations, knowledge and so forth) it will be dissimilar to the physical space of things, people and places, to which we are evolutionarily adapted. It is a space in continuous dynamics, updating its forms, a relational mobile space based on communication and virtual representations. Such characteristics will challenge the human evolutionary ability to orient and locate in "reality."

This is so, because the fundamental importance of space is not limited to the physical aspect of the human body. It is at the origin of human subjectivity, too. Primary and even deeper beneath of the self-consciousness (or transcendental ego) the subjectivity is based on the relation to the environment, perception and movements Kimura (2000). Any living being is formed by reactive modification of the inner world and consequently to unceasing modifications in the external world. Subjectivity as such is made possible by the permanent interconnection between the living organism and the envi-

ronment. This subjectivity, basically, "unrelated to consciousness, is what we call "self" (*jiko*), a principle of the connection which is established between the individual and the world and which is to be thought of as an extension of the physiological subject." (Kimura, 2000, p. 85, my translation) The conscious living presumes incessantly recognitions of the perceived objects as a given reality. But it also relates with them preconsciously, prior to recognizing them as real. This unconscious virtual perceptions of the things relentlessly emerges contingently in the conscience. The process is made possible by the pre-personal bodily self when they become individuated through actualization, adding the key input for realizing the sensation of actuality at any given moment of life. This phenomenological process is more fundamental and parallel with cognitive operation of consciousness and able to accommodate any type of space as being real, i.e. being experience as "given real space." It also opens up the possibility for different forms of transcendental conditions for the possibility of things. Likewise, it becomes more significant for the 4th Space reality where the relations between things are overpassed and lessened by the relations between persons, virtual persons, and artificial "persons" intelligences.

The seemingly familiar physical space is, in fact, very difficult to be described, as long as it is both reflected and constructed as psychological space. What people believe that they understand or experience by physical space is already a psychological construct, and usually is confounded with its abstract geometrical representation, i.e. a classical "category mistake" (Ryle, 1938). What we "thought" of being physical space is ontological and epistemological hybridization of human bio-physiology and physical reality. It is one of the multiple entities that form the psychological space. And the task of describing psychological space is almost impossible at this moment in spite of its apparent familiarity. We don't have a suitable language system able to represent and convey the complex interplay of mental and physical elements, we can only asses a very limited range of spatial phenomena with available measuring instruments or procedures of spatial behaviors and "we lack a construct which accounts for both our awareness of the relational distribution of things and our use of this awareness to solve problems" (Eliot, 1987, p. 6).

At the same time, it is certain that there is nothing in human experience that can remain purely abstract as there are ideal abstract concepts from Mathematics or Physics, for example. People can think the abstract, but they cannot live (in) it. Human thinking appeals to the concrete that is mostly based on representations and images. In addition, people cannot think and feel value-free or emotionless. Any experienced reality is appropriated by their lived experience. The living physical space becomes place ("own space"), due to its familiar psychological dimension, which implies orientations, repeated actions, organization of representations, utility and memory.

> "Physically, a place is a space which is invested with understandings of behavioural appropriateness, cultural expectations, and so forth. We are located in 'space,' but we act in 'place.' Furthermore, 'places' are spaces that are valued. The distinction is rather like that between a 'house' and a 'home;' a house might keep out the wind and the rain, but a home is where we live." (Harrison, Dourish, 1996, p. 69)

In order to understand the notions of space and place, the current mainstream paradigm of psychology should incorporate elements for phenomenological approach. Otherwise, it provides only an illusion or, at least, an insufficient knowledge of what is space for a human being, an essential element of Self formation.

## THE 4th SPACE: SPACE OR PLACE?

It has become clear now why the concept of the 4th Space seems so natural, but so hard to grasp. Its origins are to be found in the Ray Oldenburg's influential book *The Great Good Place* (1989), in the domain of communities building. This is the first trace in its conceptualization. The 4th Place (Space) is depicted as a heavenly communicational and relational place that fosters and grows human communication, relationships, works, facilitating access to and exchange of knowledge etc.

> "The fourth place blurs the frontier, within the same space, of the first (home), second (work), and third place making the space, a place in itself. The function of the 4th Space is to foster networking, to promote mingling, and to favor collaboration, face-to-face interactions, and the exchange of tacit knowledge." (Morrison, 2019, p. 448)

This conception bestows space with human attributes and unavoidable changes it into more than a simple space, a place.

The distinction between space and place is a very interesting and complex one. Place is an anthropomorphic space. Unlike space, the place has meaning, history, implies knowledge, attitudes and feelings. We can speak about a sense of place, but not about a sense of space (Tuan, 1979, p. 421). The 4th Space, as computer-supported cooperative work (CSCW)

> "is rooted in sets of mutually-held, and mutually available, cultural understandings about behaviour and action. In contrast to 'space,' we call this a sense of 'place.' Our principle is: 'Space is the opportunity; place is the understood reality'." (Harrison, Dourish, 1996, p. 67)

Hence, despite our endeavor to conceptualize and understand the 4th Space in itself, in real experience it will never exist separately from how it is felt.

Any space is impersonal, the place has identity. Space means order, place means familiarity. The space is common, neutral and imposed, and lack of any sense of belonging. A place, formed by actual living or by any other significant form of contact, is always unique, personal and created. The space is abstract, the place is lived. In space things are unrelated and indistinct, while the place involves connectedness and distinction. In any spatial located human interactions, from physical intercourse to computer mediated communication or cooperative work, the space is made by interconnection and communication. Its "ontic" multi-dimensionality (the three physical dimensions, parallel virtual space, or multiple, in extended reality) forms its appearance: an empty condition of possibility. The 4th Space is formed in actual intercourse of people, when its condition of possibility is converted in a communication place, because the space itself is meaningless and has no proprieties useful as such for human actions.

In addition, the social and cultural spaces subtly interweaved within the psychological sense of space and make it more difficult to understand. Any organism "takes its place within, toward, against an environment." Humans, in addition, once they achieve humanity, occupy their place in a world defined by roles, responsibilities and institutions. "All his life, his existence as a person is constituted by such 'place-taking'." (Grene, 1968, p. 173).

The mental spaces differ from individual to individual according to personal experience. And every culture draws different mental topographical spaces (Hall, 1966). Moreover, people are also actively "position takers."

> "The most characteristic thing people do as persons is to take positions or stances toward the elements of their experience. People take positions on events, others, issues, policies, attributes, actions, themselves, and on the positions they take. People take positions on concrete aspects of an immediate situation, on life as a whole, and on the cosmos." (Cochran, 1985, p. vii)

The continuous virtualization of the social and communicational space, technologically mediated by digital and augmented reality, amplifies all the aspects of psychological space. "The construct of psychological space refers to such an amalgam of physical capacities, mental processes, learned skills, forms of representation, and dimensions of thought at different levels of awareness for different tasks in changing surroundings." (Eliot, 1987, p. 6). It includes has some important characteristics relevant to the future development of human psychology in its dialectical process of human-living environment co-creation. Beside its common characteristics as mental basis for a set of behaviors at different levels of knowing, multimodal form of representation, a form of behavior related to the awareness of the limits and position in different environments, psychological space is also: a pervasive cognitive phenomenon, a form of symbolic processing, an expression of intelligence intricate in the higher and complex knowledge and responses to spa-

tial tasks, implied in daily activities, and even a possible dimension in many kinds of thought (Eliot, 1987, p. 6).

## THE HYBRID ONTOLOGY OF THE 4TH SPACE

So, what is it, a space or a place? The term "4th Space" instead of "Place" is supported as an analogy with physical space, but only if it is conceived as an ontological reality where things are located and processes take place and because they can be "represented," localized, and become visible and portrayed through three axis: place, medium and time (Hardegger, 2021). In this framework, the "Place-Axis" represents the grounded connection or anchor into physical of the real word place, the localization of the people immersed in the virtual cyberspace. The "Medium-Axis" represents technological supporting infrastructure, hardware and software. The "Time-Axis" represents the timeframe coordinate of user presence within the 4th Space, where the physical time distinction of synchronic/asynchronic interactions dissolves. All three dimensions form a new reality in which their correspondent from the physical reality is transfigured. In the 4th cyberspace, the physical space location of the user intertwines with its location in the digital/virtual space. The medium is both the interface and the sensorial constitution of the user.

This image is useful for theorizing on the 4th Space as epistemological entity. It provides a suitable framework for organizing the research both from different disciplines and interdisciplinary approach. However, it can still uphold a deceptive image of what cyberspace reality is by perpetuating an image of it as a preexisting territory and a metrical space. Its medium-axis component presupposes already the engineered devices and mathematical algorithms and this reduces the 4th Space to "a set of objects and rules of interaction." Its ready-made nature, "waiting to be filled" and the topographical description induces a false and apparently stark division between "real" and "virtual" worlds which is simple not true" (Mihalache, 2002).

The essence of the 4th Space is communication. In the physical space, things, states and processes are interpretations, signs, landmarks and so forth. But the 4th Space is in itself no other thing than information and communication made possible by digital virtualization. The process is similar to the cosmogenesis where the universe appears *with* space and time and not *in* space and time (the cosmic metric expansion of space-time). In the case of the 4th Space we cannot talk about a space filled *with* information, but about a space *of* information. Its very nature is information and communication.

> "The structure of cyberspace represents a hierarchy-based system of technical and semantic layers (physical, logical, information, and human) that are heavily linked to each other. The most important goods in this space are information, which is used by people, thus creating their new living space." (Gálik, Tolnaiová, 2019)

The 4th Space is one of the multiple types of culture-generated spaces as are the economic or social spaces and cannot be reduced neither to the mental space or external space. It is rather an (collective) extension of psychological space (Suler, 2016). The 4th Space completes (and at the same time modifies) Karl R. Popper's (1968) idea of three worlds, filling the point of interaction between World 2 (the realm of natural states and processes of ICTs devices) with World 3 (the objective realm of the "products of thought"). Reality, subjectivity, and representation become a triune composite. The field of reality (the physical world), the field of representation (the virtual reality) and the field of subjectivity (the individual) merged in a continuum subject-object-context of onlife living. The tension between connectedness and distinction dissolves. Technological progress made possible the extensions and substantiations of human imagination.

## 4TH SPACE AS TECHNOLOGICAL EXTENDED REALITY (XR)

The 4th Space ontology is hard to be conceived (represented or imagined) because it has a very heterogeneous, complex and dynamic nature. It is more than people using and communicating on their laptops and smartphones. The 4th Space has an insidious ubiquity and from this the difficulties to understand what its reality is. In XR, we find the second trace for the 4th Space conceptualization. Here, all the human functions, abilities, potentialities are transformed, magnified, extended, augmented and mediated. CSCW is accomplished as the 4th Space in the framework of *ubiquitous computing* (or *pervasive ambient computing*). It comprises an enmeshing of the properties of various types of intelligent devices that unfold the physical world in an ameliorated virtual reality of communication, interrelations and perceptions permeated with huge amounts of collective knowledge. However, even the nature *of* pervasive ambient computing is difficult to grasp in a clear-cut image. Stefan Poslad (2009) observes that the definitions and descriptions of pervasive computing are overlapping and puzzling. Therefore, he proposes that it can be better understood by a taxonomic systematization of its proprieties: *distributed system properties*, *implicit human device interaction*, *context aware*, *autonomy*, and *intelligent system properties*.

The proprieties of the 4th Space result from the synergy of ubiquitous computing individual devices, programs, networks and multiple agent sys-

tems. Like the less practical endeavor to define ubiquitous computing, the 4th Space reality can be better understood as the result of the dynamic system of myriad of artificial (intelligent) entities with mixed and diverse proprieties. The *distributed system properties* can be universal (seamless or heterogeneous), networked, synchronized (coordinated), open (transparent or virtual), or mobile (nomadic). The *implicit human device interaction* covers proprieties as non-intrusive (hidden, invisible or calm computing), tangible (natural), anticipatory (speculative or proactive), affective (emotive), user aware, post human, or having a sense of presence immersed (virtual or mediated reality). According to their capacity to be *context aware*, the systems can be sentient (unique, localized or situated), adaptive (active context aware), person aware (user aware, personalized or tailored), environment aware (context aware or physical context aware), or ICT awareness. From the point of *autonomous system properties* they can be automatic, embedded (encapsulated or embodied), resource constrained, untethered (amorphous), autonomic (self-managing or self-star), or emergent (self-organizing).

The *artificial intelligence* can be individually or collectively distributed. The individual intelligent systems can be reactive (reflex), model based (rule/policy based or logic/reasoning), goal oriented (planned or proactive), utility based (game or theoretic), or learning (adaptive). Multiple intelligent systems (collective or social intelligence) can be cooperative (collaborative or benevolent), competitive (self-interested, antagonistic or adversarial), orchestrated (choreographed or mediated), task sharing (communal, shared meaning or knowledge), speech act based (intentional or mentalistic) or emergent (Poslad, 2009, pp. 17–22).

The 4th Space arises from the synergy of the diverse proprieties of various devices and systems and creates a novel dimension for human existence. The ubiquitous computing does not only create a space but, by customized adaptation to user preference, builds a new place.

## VIRTUALIZATION AS CULTURAL EVOLUTIONARY PROCESS

The third trace for understanding the nature of the 4th Space is its virtuality. Virtuality is another key element for understanding the 4th Space, but more in the sense of communicative structures, not in that of artificial reality. In this sense, the 4th Space history and genesis is longer than many may think. The communicative virtual communities existed long time before the Internet (Stone, 1991, pp. 94–95). Its origin can be found in the first *virtual communities* created by the invention and dissemination of the texts. Textual virtual communities fostered a community of like-minded persons

through intellectual interchange mediated by books. It has endured today in the network of scholarly publications. The next stage of building virtual community was the rise and spread of *mass media*. Early electronic virtual communities (made possible by radio and television) connected people synchronously and made it possible for many people to be transposed and being "present" in the same informational location from remotely physical space. The next stage, the *era of information technology* of the World Wide Web represents the true birth of active virtual communities. The commonality fostered by passive reading community of knowledge became active and interactive, "a participatory social practice in which the actions of the reader have consequences in the world of the dream or the book." In the third stage,

> "... the older metaphor of reading is undergoing a transformation in a textual space that is consensual, interactive, and haptic. [...] The boundaries between the social and the natural and between biology and technology are beginning to take on the generous permeability that characterizes communal space in the fourth epoch." (Stone, 1991, p. 95)

The last epoch, that of "real" virtual reality, is thought to be fully accomplished in a space akin to Gibsonian Matrix (Gibson, 1984). In the maturated 4th Space the real and virtual will merge in an inter-psychic network of a communicational field that extends the physical reality in a new one.

In contradistinction to virtual reality, which evokes artificiality, the bogus or constructed character, the non-real, and hence emphasizing reality, the cyberspace highlights inclusively the actual place. The 4th Space is rooted in cyberspace and even opposed to virtual reality because it "does not rely mostly on a deception of senses to create the illusion of an integral realism," but it is a space for computer mediated communications (Holmes, 1997, p. 234). The cyberspace exists only as a communicative function of its inhabitants. A single person exists only in virtual reality, but not in cyberspace, because the critical element cyberspace is community (Ostwald, 1997; Benedikt, 1991).

Due to its particular characteristics, communicational nature, pervasive computing, and XR, it is more than a reasonable expectation that the new onlife living to have a meaningful effect on the basic sensory-psychological coordinates of space and time on which personal identity is initiated. The prolonged onlife living (Floridi, 2015) within the 4th Space will definitely have dramatic effects on the personal identity construction, as long as living "in reality" is based on cognitive operations and bodily actions.

## LIVING IN 4th SPACE EFFECTS ON PERSONAL IDENTITY

The difficulties in crystalizing a coherent personal identity in modern world are illustrated by the big challenge in understanding the living environment, i.e. the dynamic enmeshing of physical, augmented and virtual reality of onlife world (Floridi, 2007). Personal identity is much deeper molded by our new "onlife" way of living, embedded in ICT's technology and digital relationships, which changes "our self-conception (who we are); our mutual interactions (how we socialise); our conception of reality (our metaphysics); and our interactions with reality (our agency)" (Floridi, 2015, p. 2).

In order to prospect how the prolonged living in 4th space can alter the personal sense of identity, we cannot rely solely on current mainstream psychology because the algorithmic structure, methodological quantitativism and underlying computational paradigm, though very useful for some epistemological goals, are of little use for catching the dynamic process of personal identity formation. The (post)modern identity needs to employ a more complex approach and conceptualizations, although not as simple to understand and impossible to be algorithmically modeled. I am referring to the rhizomatous conceptualization and the phenomenological approach.

## THE RHIZOME AND RHIZOMIALITY

During these times of technological changes and challenges, understanding the systemic factors contributing to the transformation of the personal and moral identity of nowadays person(s) is a very difficult task because the formation is not a linear and homogenous process. The difference in Self-formations becomes increased because the changes brought by onlife living in personal identity are facilitated both by rhizomatous characteristics of the self (Baldwin, Greason, Hill, 2018) and by the "rhizomality" (Deleuze, Guattari, 2005; Deleuze, Guattari, 2005) of the contemporary (digital) social space (Kalantzis-Cope, Gherab-Martín, 2010). In my view, the rhizomatous pattern of conceptualization is a more suitable framework for grasping the formation of the 4th Space inhabitants' personal identity.

A rhizome is an onto-epistemological model of heterogeneous multiplicity in which the organization of the elements does not follow a subordination line. "A rhizome is not amenable to any structural or generative model. It is a stranger to any idea of genetic axis or deep structure" (Deleuze, Guattari, 2005, p. 8). Connection, heterogeneity, multiplicity, asignifying rupture, cartography and decalcomania makes rhizomatous representation the best candidate for understanding the dynamic of identity and moral development of digital natives (Deleuze, Guattari, 2005). The dialogical, relational and

hybrid nature of interactions in the 4th Space are reflected in the construction of the personal identity as narrative ISelf (Popoveniuc, 2017). We are permanently connected in physical and virtual life with others by knowledge, interpreting and applying rules, norms, and different values. Living onlife means a heterogeneous life where the virtual and physical realms dialectically reinforce each other. In the 4th Space, physical reality and extended reality intermingles in a uniform hyperreality where human and artificial intelligence fuse (Tiffin, Terashima, 2005). The seamless fusion of the physical and the virtual and of the digital semantics and human cognition is facilitated by the rhizomatous proprieties of the mind and the cyberspace.

> "Any point of a rhizome can be connected to anything other, and must be. [...] A rhizome ceaselessly establishes connections between semiotic chains, organizations of power, and circumstances relative to the arts, sciences, and social struggles. A semiotic chain is like a tuber agglomerating very diverse acts, not only linguistic, but also perceptive, mimetic, gestural, and cognitive: there is no language in itself, nor are there any linguistic universals, only a throng of dialects, patois, slangs, and specialized languages. There is no ideal speaker-listener, any more than there is a homogeneous linguistic community." (Deleuze Guattari, 2005, p. 7)

The very nature of the 4th Space embodies a real multiplicity of states, realities, knowledge and values. There is no fixed point, be it ontological, axiological or semantic, to support the complete division of subject and object. Individual humans become only parts of the multiplicities of semantics, virtual and cognitive realities in the vast networks of multi-agent system. In the rhizomatous paradigm, "the multiple is effectively treated as a substantive, 'multiplicity,' that it ceases to have any relation to the One as subject or object, natural or spiritual reality, image and world" (Deleuze, Guattari, 2005, p. 8).

And there is another reason why the rhizome model fits better to understand the contemporary reality. The fast pace of digital transformation, augmentations and amelioration of humans, environment and culture puts us on the progression path toward "trans" and "post" humanity. It is consonant with the transhuman state of modern man which is significantly technological (bio)ameliorated (by taking pills, wearing glasses, cognitively extended capabilities etc.). The prefix "trans-" acquires and expresses the dynamic processual state of "meta-" being "in the middle of" but at the same time "beyond" and "between" (Sorgner, 2020). The 4th Space expresses both, in its ontological reality, the "trans-" or "meta-" space of reality and, in its sociological nature, the "trans-" or "meta-" sociality. Similarly, "a rhizome has no beginning or end; it is always in the middle, between things, interbeing, intermezzo" (Deleuze, Guattari, 2005, p. 25).

The consequences of living and thinking in such rhizomatous environment are far more thoughtful. The virtualization of *technics* and technology in the forms of digital and XR, as a process of reversal exteriorization of autopoietic human cognition into the environment, accomplished the "closure of the cortical evolution of the human" and opens the way for *the pursuit of the evolution of the living by other means than life* (Stiegler, 1998, p. 135).

## A PHENOMENOLOGICAL APPROACH OF THE 4TH SPACE

Any living beings maintains the continuous contact with its environment in order to survive. This is also true for the groups of any species. The contact can be represented by physical surface of the body or by any of its surrogate a cane, car body, computer screen, VR glasses etc.

> "Our self-awareness, our experience of the 'I', is based on the *Tatsache*—an expression literally meaning 'the thing of acting'—that is, things or objects perceived in the outer world or imaged in the inner world, are associated with a quality of actuality, *Wirklichkeit*, which endows the perceived objects or imaged with a sense of reality, and the perceiving or imaging subject of experience, an 'I'." (Kimura, 2008; cf. Kimura, 1963, p. 391)

When this phenomenological character of qualia is disturbed, the pathological state of personalization can occur. The sense of reality of the external world and the sense of [its] own existence results from the same processes.

The individual subjectivity, understood as a "continuous discontinuity" relation with the surrounding environment, is doubled by collective subjectivity of appurtenance group (and groups). "This fusion of the human I with the auto-affection[1] of life in general, constitutes a necessary condition of a healthy mental state" (Kimura, 2001, p. 336).

Kimura's psychiatric conception hypothesizes a certain fundamental dissociation between individual existence and existence as a constituent of the species, as the fundamental process affecting the I-sense in the schizophrenia. In the context of our discussion on the personal identity formation within the 4th Space, the "schizophrenia" should not be taken in its psychiatric sense, but as a model or label for any other different or alternative mode of the sense of Self formation and, hence, of constituting the self-identity. The continuous mediated and intermediated contact with exteriority can elevate a higher level of self-reflection and self-reference which entails a deep uncertainty about "the I-ness of the self or the selfness of the I." This basic dis-

[1] *The affection of the self by the self*, the purely immanent, unmediated self-affecting relation is a key term in Henry's phenomenological conception. "Auto-affection is the internal structure of the essence whose property is that of receiving itself." (Henry, 1973, p. 236). "Auto-affection thus describes how life affects itself and how it receives its affect" (see Sackin-Poll, 2019).

turbance of "discordance between individual subjectivity and the collective subjectivity to which one actually belongs" can be "a potentiality inherent in all human beings" (Kimura, 2001). But his disturbance can be conceived, not necessarily as "discordance," but as an expression of the plasticity of the I-Self that can "accord" in many ways with the collective subjectivity. The subjectivity can be differently constructed, alike its space(s) as condition of its possibility. Onlife living in the 4th Space can thicken the borders of I subjectivity and group subjectivity due to the dissolution of Otherness in virtual avatars and mediated homogenous reality. Or, on the contrary, it can facilitate a more harmonious integration of the individual and the collective self.

No matter if we are optimistic or pessimistic about the human integration and amelioration with and within technological progress, the onlife virtual-real living will have profound effects on the basic mechanism of constructing personal identity and self and the sense of the "I" as personal subject of experience and action. The consequences of prolonged life in the techno-environment will be, definitely, more complex, deep, and diverse. The growth in consistency of the 4th Space implies an increasing dependency on technological created reality and increasing living in blended reality that can blur extensively the borders of "here and now" (Waterworth, Hoshi, 2016). The effects of prolonged living and extended interactions with virtual and augmented environments does not exclude even the possibility to "lead to more fundamental changes, not only on a psychological, but also on a biological level" (Madary, Metzinger, 2016, p. 4).

We see now that the 4th Space is neither "a new stage of *etherealization* of the world we live in," nor "a new stage in the concretization of the world we dream and think in," but another genuine "venue for the consciousness itself" (Benedikt, 1991, p. 124), a new existential dimension of man (Gálik, Tolnaiová, 2019). This claim can sound bombastic or unusual within the mechanistic and positivist epistemological paradigm of the mainstream sciences. But it becomes not only natural, but self-evident, once it is set in the required new epistemological paradigm of ICT described above. The homogeneity between Self/personal identity and the 4th Space is more salient when we take into account their similar process of autopoietic genesis.

## Identity as autopoiesis

The living beings and the conscious beings form themselves in a process of autopoiesis. An autopoietic system is a dynamic self-sustaining

> "network of processes of production (transformation and destruction) of components that produce the components which: (i) through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them; and (ii) constitute it [...] as a con-

> crete unity in the space in which they (the component) exist by specifying the topological domain of its realization as such a network." (Maturana, Varela, 1980, p. 79)

In short, the autopoietic system continuously self-produces both itself, as a network of the production processes, and its own space, its boundaries. As the Universe appeared *with* its physical time and space, the self-poietic systems produce themselves *with* their own space (and time).

> "Life is realized in a space of its own to inhabit it, but it does not form itself in a space designated by an observer, everything is integrated into the organism itself. The formation of oneself is at the same time, so to speak, the formation of a space which is proper to it." (Kawamoto, 2011, p. 352)

They form for themselves by virtue of their own operation that also provides their own domain of existence. This does not exclude the influence of the outer world, because the self-poietic systems do not appear in void but within an environment to which they adapt accordingly. The formation of the Self through auto-genesis is akin to that, but at a different level of virtualization. The 4th Space is just another niche for the evolution of life as a cognitive and information autopoietic system. The intimacy of Living Space and Self Identity genesis is revealed by their common mechanism of formation, i.e. the autopoiesis.

In Hideo Kawamoto's conception of self-poietic systems, the self is a self-poietic system that emerges through its own actions and movements. It creates the borders of the self and non-self. It self-produces itself through bodily action "by doing something" and less by cognitive sense of "knowing something." In this process one produces reality through action, persons, and things and the fittest and surviving of the individual system depends on the coherence between its actions and the characteristics of the environment and things. In the case of 4th Space cyberreality, the "coherence" between actions and the environment, due to its fundamental informational and relational nature, becomes fluid. The plasticity of virtual environment and things will make the Self more fluid and less stable because the relation between one's actions and the environment characteristics is less stable. The required felt constraint to adjust one's action to reality is weakened. From here, the increase freedom and diversity for self-poiesis of personal self and identity within the 4th Space.

Luciano Floridi (2011) makes an interesting "theory transfer" of autopoietic systems concept into the philosophy of information that is extremely illustrative for understanding the particularity of the genesis of the Self within the 4th Space as informational and communicative environment. In his theory of the Relational-self, he identifies a unique mechanism of encapsulation, detachment and internal auto-organization at all three levels of

progression of life, i.e. from corporeal, through cognitive, to the consciousness as a progressive process of virtualisation. Any auto-organizing system develops a membrane for protecting its structural integrity against the surrounding environment. The conscious personal identity is triune (three-leveled) entity. The *corporeal membrane* or physical membrane functions as a hardwired boundary between the inside (body) and the outside (the environment). The *cognitive membrane* is a semi-hardwired (configurable) boundary between the cognitive system and its environment (body becoming an interface) that further isolates and controls the organism from its surroundings, using data processing and communication for maintaining the integrity of the organism. *The consciousness membrane* is a soft-wired (programmable) boundary between self, conscious mind or I (inside) and body (outside environment). The homeostasis at each of the three levels implies different types of bonds and orientations: physical stability of the living system is based on chemical bonds and orientations, the integrity of internal data within the system is assured by mutual information, that is the (measure of) the interdependence of data, and their codification in memory, and language, while the self-programing within the cognitive system is provided by semantics. "The self emerges as a break with nature, not as a super connection with it" (Floridi, 2011, p. 560). The virtualization of the structures of the Self imply increasing autonomy up to the world of meanings and interpretations. This cannot be realized individually, but only in common through language, culture and social interactions. Floridi's account highlights the inherently free relational Self and not as the traditional metaphysical conceptions, only the rational disembodied self. The onlife in the 4th Space reflects accurately this situation. However, this perspective misses two important questions: the required compatibility between semantic elements and reality and the psychological aspects of the Self's formation, which is not only an informational system. It only moves the question of compatibility between individual cognitive structures, which interprets the physical processes and states, to the collective level of the viability of cultural semantic tools.

Niklas Luhmann's (1988) theory transfer of autopoietic systems on social systems theory can fill this gap. Its perspective is also extremely relevant for Self formation within the 4th Space framework because the intimacy between Self formation and the nature of the 4th Space is simply striking. The social systems are also self-organizing and self-reproducing systems, unlike the physical environment, hence they are exclusively communications and not ensembles of individuals, roles, acts, and/or interactions between them. (For the validity and utility of such theory transfer see Cadenas, Arnold, 2015). The transdisciplinary applicability of the theory of autopoietic systems moves the key element on the self-referentiality of the autopoietic systems. Autopoietic organization of the social system is built on communi-

cation, while psychic autopoietic system on consciousness as modes of meaning-based reproduction.

## IDENTITY IN THE 4TH SPACE

The complexity of modern environment and cultural settings in which people are living challenges the consistency and substantiality of personal identity. In simpler societies of the past, the unity and coherence of personal identity, and hence personal morality, was ensured by relatively homogenous living conditions, shared moral norms, knowledge, ideology, religious beliefs and uniform conduct of the others from the same society. In modern societies, the individual is exposed, from its infancy, to a lot of heterogeneous practices, various ways of conduct, conceptions, knowledge, values and so forth. This miscellaneous cultural and social environment diminishes the relevance and ability of the external social environment to support the consistency and substantiality of personal identity. As consequence, nowadays "a person's identity is not to be found in behavior, nor—important though this is—in the reactions of others, but in the capacity to keep a particular narrative going" (Giddens, 1991, p. 54). This "internalized integrative narrations of the personal past, present, and future" provides unity, purpose and meaning in one's life (McAdams, 1996).

In the 4th Space, the difference between virtual and real stimuli, semantics and representations becomes blurred and shattered. The stable, rigid and, steady framework for separating or cutting across a single structure by over-signifying breaks becomes impossible. The meaning, interpretations, hermeneutics of real and reality, of relations and interactions become malleable and flexible. The narrative path of self-identity can be changed, transformed, reinitiated, converted, resignified from the beginning with materials and experiences borrowed from physical reality and face-to-face interactions or from virtual, mediated ones from virtual and extended reality. Nothing more auspicious for the growth of rhizomatous autopoietic Self. "A rhizome may be broken, shattered at a given spot, but it will start up again on one of its old lines, or on new lines" (Deleuze, Guattari, 2005, p. 9).

In the 4th Space, personal identity, in general, and its core characteristics, moral identity, in particular, grow into a rhizomatous interaction of private space values and moral, professional space values, public space values, digital (augmented) space values. Living in a continuously shifting and mixed virtual-physical place entails a completely resignification of fundamental values in general, and the professional and public values in particular.

> "The distinction between public and private will probably need to be reconceptualised, because frameworks based on physical boundaries (the ever per-

> vasive analogy of trespassing) and possession (the equally pervasive analogies of ownership and theft) are outdated conceptual modules, insofar as they are linked to a modern or 'Newtonian' metaphysics based on inert things and mechanical interactions." (Floridi, 2015, p. 22)

At the same time, personal identity is based on self-narration that is based, it its turn, on memory. The neurophysiology of memory already confirmed the different neurological and cognitive mechanisms of long-term memory and short-term memory (Norris, 2017; Nee, Berman, Moore, Jonides, 2008). The personal identity is related to long-term memory, but ceaselessly mediated by short-term memory, which processes the information prior to be stored in long-term memory each and every time when it is consciously and unconsciously accessed. Identity is therefore fostered by the permanent interactions between the continuously interpretations of experience, which is constantly processed in short-term memory, and the fixed conceptual and identity arrangements treasured in the long-term memory. The long-term memory offers the core basis for personal identity, but all the images, concepts, memories, autobiographic episodes, states, ideals values stored are incessantly slightly changed, eroded, transformed, and molded by the permanent working memory in daily experience, like tide's waves imperceptible change in time the shoreline. Their working mechanisms are quite dissimilar.

> "The difference between them is not simply quantitative: short-term memory is of the rhizome or diagram type, and long-term memory is arborescent and centralized (imprint, engram, tracing, or photograph). Short-term memory is in no way subject to a law of contiguity or immediacy to its object; it can act at a distance, come or return a long time after, but always under conditions of discontinuity, rupture, and multiplicity." (Deleuze, Guattari, 2005, p. 16)

The personal experiences are also processed by employing the concepts and information from long-term memory.

> "Short-term memory includes forgetting as a process; it merges not with the instant but instead with the nervous, temporal, and collective rhizome. Long-term memory (family, race, society, or civilization) traces and translates, but what it translates continues to act in it, from a distance, offbeat, in an 'untimely' way, not instantaneously." (Deleuse, Guattari, 2005, p. 16)

Living onlife in the 4th Space implies a significant increase in the short-term memory role for personal identity. The prevalence of the iconic increases the importance of sensory memory for constructing the self-image. The individual is exposed too more and diverse relevant models, values, and experiences. The personal identity seashore is more deeply and relentlessly changed by the broken waters (Camina, Güell, 2017).

The organization of the 4th Space, as open and self-created communicational and informational space, provides a different type of knowledge accumulation and progress. In Eco's (2014)metaphors, its rhizomatous structure "encyclopedically" constructs information, by *mapping* knowledge as a network of interlinked relationships, and not as dictionary, as a closed system of semantic and informational *trancing*. The fixed and absolute knowledge is lost, in the favor of freedom to pursue and develop an infinity of new connections and meanings. So, the search for origin and reconstruction of some linear development becomes less relevant for the personal identity's self-narrative. The substantiated multiplicity of personal identity's bricks reveals a plural origin, heterogeneous incentives and a variety of sources, sometimes contradictory, segmented, with suddenly broken paths of development, reinitiated and re-signified. "Within the rhizome, thinking means feeling one's way, in other words, by *conjecture*." (Eco, 2014, p. 55)

Self-identity can be more free and reflexive than ever. People have access to information that allows them to shape their personal identity in unimaginable and various ways, to reflect on the causes and intended and unintended consequences of their own action and especially to have many cultural alternative and axiological frameworks for interpreting their behavior. Out of this liberation from the stable, structured model of personal identity comes also its curse. This freedom to choose who we are and how we understand ourselves and our own conduct is practically equivalent with the condemnation for full responsibility on our deeds, as Jean-Paul Sartre (1943) remarked. And this is an anguishing task for anyone. Moreover, the increasing capability for individual self-poiesis is threatened by the dispersion of the sense of self in the multiplicity of interpretation. The peril of losing the human wisdom in knowledge and knowledge in information, envisaged by T. S. Eliot (1934).

## THE MORALITY IN THE 4th SPACE

The emergence of the 4th Space reveals the importance of "ethical infrastructure" or infraethics "as a first-order framework of implicit expectations, attitudes, and practices that *can* facilitate and promote morally good decisions and actions. Examples include trust, respect, reliability, privacy, transparency, freedom of expression, openness, fair competition, and so forth" (Floridi, 2017; Floridi, 2012, pp. 738–739). I will not discuss here Floridi's (2017) list of "inframoral" entities neither in its completeness (why, for example, the belief is not listed among first in importance?), nor as a possibility of existence, even theoretical, at the collective level of a society made up of less moral individuals. If what it is called the "infraethical" structure can be made possible by the average moral level of its bearers. For

me, at least some elements are, in fact, parts of an "ultraethics" of that society, i.e. the networking ethical aggregate of myriads of micro-moral relations. For our subject it is significant how the ICT designed society affects, what Floridi call the "infraethics" of a given society.

In the 4th Space world, trust, reliability, transparency etc. become distributed between multi-agent systems (Wooldridge, 2009). People must put their trust not only on other people, but also on artificial systems, software, programs, and multiagent (human and artificial altogether) systems. Because the later are more predictable, do not have agency, occult interests or being considered objective experts, they become to be considered, sometimes, more reliable than living fellows. As long as the people's actions, existence, decision-making are increasingly intertwined in long chains of dependence on ITC infrastructure and program designs, networks of coworkers, multi-agent systems and intelligent systems expertise the moral agency diffuses in a *distributed morality* (Floridi, 2012). Ethical agency-assigning, accounting and assuming ethical responsibility for agents—becomes abstruse in the process of melting frames of physical, artificial, legal, psychological, public and private spatial borders within the 4th Space.

At the same time, the dispersion of responsibility (similar with the well-known phenomenon of diffusion of responsibility (APA, n.d.) is very likely to become more common and prevailing within the 4th Space. Here, all aspects endorsing the diffusion of responsibility are amplified. *Anonymity* is more prevalent in the 4th Space, even only under its subjective, if not objective, sense of it (Postmes, Spears, Sakhel, de Groot, 2001). People are in a paradoxical state within the 4th Space. They are simultaneously interconnected with many others more than ever, but they are also atomized because of the computer mediated relations with the others. Such situation increases additionally the diffusion of responsibility brought by the *division of labor*. Being more individualized as ever in working on their tasks, people can lose the general sight of organizations as a whole, that they are part of a general syncretic-synchronic system, and limit themselves to narrow tasks assigned for them (Baumeister, 2015). The differing and conflicting types of *expertise*, human versus artificial systems, and the failure to capture important information diminish the level of responsibility and accountability for personal deeds and contributions felt by individuals (Wegner, 1986). Not to mention the *group size*, indefinitely in the 4th Space settings, that is indicated as one of the key factors for decreasing the sense of individual responsibility (Latané, Nida, 1981; Rowan, Kan, Frick, Cauffman, 2021).

The pervasive infrastructure of AI system supporting the 4th Space contributes to this phenomenon of distributed morality that is the "macroscopic and growing phenomenon of global moral actions and non-individual responsibilities, resulting from the 'invisible hand' of systemic interactions among multiagent systems (comprising several agents, not all necessarily

human) at a local level" (Floridi, Sanders, 2004). The sense of responsibility is easily weakened in such "moral crumple zone" (Elish, 2019), where the consequences of the actions of a human actor depends on the behavior of an automated or autonomous system on which he/she has a limited control.

The hybrid onlife living can also have very perverse affects due to the impossibility of mind and body to differentiate virtual experiences from real ones. A person can live as a virtual avatar through fulfilling his/her unsatisfied needs, expressing his/her darkest desires and thus liberating himself/herself from its frustrations. At same time, it can do this dissociated from the responsibility of his/her own deeds, because can create the world's rules and norms. "Rather than simply being influenced by technological features, users have intentional and purposeful control over VR stories" (Shin, 2018). The 4th Space may be the ultimate games of reality where people can do what they want with less or without the fear of consequences. It is hard to believe that moral mechanisms, fostered in thousands of hundreds of years of evolution, can act efficiently as a control mechanism, as long as their basic affective triggers are worn and diminished by the lack of physical interactions.

From here the need for large scale, innovative and open-minded prospective psychology researches. The time variable and real "virtuality" (sic!) is essential for understanding the effects of living in virtual reality. Evidence from short-term and particular aspects in laboratory settings seems to entail a more optimistic perspective on application of VR as ultimately "emphatic machine". In my view, the results are more "wishful conclusions" on the potential and local effects of therapeutically interventions using VR, while the many ethical concerns regarding its legitimacy are overwhelming (Rueda, Lara, 2020). The ecological validity of this knowledge on VR in general and effects of the prolonged and substantial life in virtual reality is very questionable (Ventura, Badenes-Ribera, Herrero, Cebolla, Galiana, Baños, 2020). On the contrary, the moral mechanisms of the human psychology have limited power even in the physical and face-to-face relations. So it is dubitable that their functioning in cyber reality will be efficient. The 4th Space relationship can have both a detrimental effect on empathy, one of the key component of moral psychology, and augmenting one or many of the dark personality traits (Kircaburun, Griffiths, 2018). The lack of physical interactions diminishes the triggers of prosocial affective mechanisms, as affective and compassionate empathy. The hybrid physical-virtual reality of the 4th Space and augmented reality settings disturb the natural mechanism of accountability and agency. "In the physical embodied world, we have no choice but to assume responsibility for our body actions. [...] The possibilities inherent in virtuality, on the other hand, may provide some people with the excuse for irresponsibility" (Turkle, 2011, p. 254). The shocking sexual assault and verbal harassment experience of the psychotherapist who con-

ducts a research on the Metaverse (Patel, 2021) is a very worrying warning about the inadequacy of moral psychology mechanisms for the life in a 4th Space reality and the perils of how dystopic virtual reality can be.

The enmeshing of what is artifactual and of what is natural, of what is psychological and virtual, blurs the distinction between reality and virtuality, between human, machine and nature. Yet, the borders of reality itself, conceptualization, normativity, identity, and relationships fade and change (with the increasing of the virtual resources in relation to the material resources). Consequently, people are challenged to manage multiple identities, including the moral ones. In a substantially digitalized society, there is place for playing, even simultaneously, several different social roles, which implies diverse sets of interchanging different moral identities. This is why the modern endeavors to decode the moral mechanisms and moral identity within the mainstream of the cognitive psychology are miss oriented.

As I have showed, the changes detected in the formation of personal identity and behavior, resulting from our moral identity/identities are facilitated both by our own rhizomatous Self and by the "rhizomality" of the contemporary (digital) social space where the virtual environment is contributing to and influencing our identity/identities and consequently our moral identity/identities. The "rhizomality" can rapidly lead to transgression, as the virtual reality is a rhizomatous space providing a possibility of multidimensional, multileveled and multipolar transgression of moral norms for any individual immersed. Why is transgression made more likely?

The rhizomatous formation of the self has freedom of plasticity and can be changed and transformed in continuous experience of the real which also is dual-sided: real and virtual. The identity became map-like, rather than trace-like.

> "What distinguishes the map from the tracing is that it is entirely oriented toward an experimentation in contact with the real. The map does not reproduce an unconscious closed in upon itself; it constructs the unconscious. [...] The map is open and connectable in all of its dimensions; it is detachable, reversible, susceptible to constant modification. It can be torn, reversed, adapted to any kind of mounting, reworked by an individual, group, or social formation." (Deleuze, Guattari, 2005, p. 12)

Being map-like the ISelf identity is more like a continuous performance rather than a more stable structure of "competence". For moral Self perspective, this implies a diminished power down to disappearance of fixed and stable deontological principles of conduct. Similarly, the utilitarian rules of decision making are too complex to be constantly followed in daily interactions. The only viable base for moral behavior remains the virtue ethics. But the virtues are attributes of character based on constant performance. It requires a very elevated and comprehensive framework for incessant ac-

commodation the personal experience in the flexible moral matrix without corrupting it. Such framework should be also collective widespread and this can be made only by a valid and trustful social institution as Science, in particular, Psychology. But the required characterological research and education implies psychodynamics, self-reflexivity and humanistic perspective. It is unfitted for stark quantitative description and, therefore, is marginalized from current mainstream psychology. This is a significant argument for a "real" revolutionary change in its paradigm (O'Donohue, Ferguson, Naugle, 2003).

However, the happy and optimistic 4th Space communality can have even a darker evolutionary path. The particular characteristics of overcrowding, abuse of virtual interactions, lack of privacy of the cyberspace environment are similar to those found at the basis of "behavioral sink" that can lead to the development of a pathological aggregation or a pathological togetherness of the inhabitants (Calhoun, 1962, p. 295). More than that—as Edmund Ramsden explains, referring to sociologist Louis Wirth—incessant aggression, frustration, interference and conflict that finally lead to "depleted social relations, personal grief and personality disorders" are consequences of an overload of social interaction. Moral decay results "not from density, but from excessive social interaction" (cf. Garnett, 2008). The 4th cyberspace co-living, with erased borders between public and private, between shared and intimacy is prone to augment sorts of similar situation.

Not to mention that I did not take into account the systemic detrimental effects resulting from economic interest and political manipulation that erode the moral stability of any given society. On the contrary, the effects of economic stark interests with its associate mentality, affecting the cohesion of society and mechanisms of power are expected to be greater within the 4th cyberspace than in the physical face-to-face society where the evolutionary acquired affective mechanism regulating moral behavior is still powerful.

It is known the increasing shift in importance from the power as physical force and material possession to power as knowledge and valid information. And that is especially important for understanding the "genealogy" of moral identity and morals. The essential nature of interconnectedness which characterizes the 4th Space makes the relational nature of power more salient and makes possible its ubiquitous presence in social networks in the highest degree. People do not poses rather that they are fostered by this power networks, which emerge from the milliard of interactions between people and people and technology (Foucault, 2006).

## Education for living in the 4th Space

Education for new (prospective) onlife ethics becomes first priority for the stability and sustainability of future society. The non-formal or formal education for science, moral-democratic competence (Lind, 2019), critical

thinking (Fasko, 1994) and for creativity and innovation support and multiply the perspectives on what is or is not a good, right, moral, enhancing our moral identity. These are the privileged cultural spaces where the desirable changes in the moral identity can take place. They stand as flexible reasonable milestones for the inevitable multiplying transgressive acts of the refusal to obey any given conditions, since the very existence of the boundary/boundaries of what is moral presupposes their violation and since reality itself became more a possibility than an actuality. A better understanding of these processes will allow universities, schools, and educational policies makers to adapt their curricula and practices, in order to fit the challenges of the digital age. "ICTs have made possible unprecedented phenomena in the construction of the self. Self-poiesis today means tinkering with the self, with still unknown and largely unassessed risks and rewards" (Floridi, 2011, p. 565). And, taking into account the rapid pace and extended technological progress, this is only the beginning. "Unfortunately, as if this were not already a gigantic task, it needs to be paralleled by the development of an equally robust ethics of self-poiesis, a new ecology of the self fully adequate to meet the demands of a healthy life spent in the infosphere" (ibidem). In the beast case, the educational polices are focusing now on the methods and forms of how to deliver the knowledge to digital natives, when the chief question is "what kind of knowledge will be required and expected when living onlife" (Floridi, 2015, p. 22).

In order to foresee and provide the suitable education, firstly the cultural base for understanding this brave new technological world must be enhanced. Science itself must be transformed from the handmaiden of technological progress into an enlightened system of understanding. The foolishly hiding behind the principle of objective knowledge doesn't work in this domain of technological progress where science makes reality, not discovers it.

## "FUTUROLOGICAL" RESEARCH DIRECTIONS

Being a social and informational autopoietic system, the 4th Space is also a meta-communicational system that can communicate about its own communications and to choose its new communications. As Geyer (2001) observed, the autopoiesis way of action has major consequences on the epistemology of these social, communicational or informational systems. While the autopoiesis and observation are distinct, the observing systems, epistemological framework for study, is also autopoietic, and thus subject to the same condition. They are reconstruct and self-modifies in the very act of observing the autopoietic system. This is an impossible conundrum for the classical logic of current epistemology, which cannot consistently accommo-

date both fundamental distinctions between autopoiesis and observation, and between external and internal (self-)observation. In any autopoietic system (social, communicational, informational) "all observations are by definition self-observations" (Geyer, 2001).

The current mainstream of scientific and philosophical enquiry on reality, i.e. cognitive sciences and analytic philosophy, is of little use, also because it misses the phenomenological aspect of the identity formation that is based on the dialectics between reality and actuality. The epistemological and academic success of these approaches is deceptive as long as they are tailored on and by the structure of cyber-reality: technological advancement, computer science, and neurophysiology. The current mainstream of psychology (cognitivism) and philosophy (analytical philosophy) are mostly tools for developing more ameliorated cyberspace for onlife living and to pave the way for a general rational (and algorithmic) cyborg intelligence. Without a humanistic perspective, psychology and philosophy become the *ancilla* of AI. Their answers and perspectives are limited to and molded after the technological and AI Weltanschauung. The human phenomenology and conscious experience are beyond their reach, together with deeper questions left unanswered by them, simply because these are beyond their legitimate horizon of interrogation. Human consciousness is a self-regulated, autopoietic informational system with characteristic semantics and states irreducible to physical substratum. The techno-naturalization of contemporary scientific paradigm can miss the essence of humanness. But this is still a small peril. The big peril is that, as the main paradigm for understanding the human being, they can have detrimental effect on its humanity. Humans can cease to see themselves "humanly" and to conceive themselves as "human" anymore. The use of cerebral implants has big chances to lead to erasing any border between what is human and what is artificial (Fukuyama, 2003). At this point the debate about naturalistic fallacy and fallacy of naturalistic fallacy become extremely relevant.

It is important to include in the mainstream psychological paradigm new methods and new theoretical models able to incorporate rhizome-like structure of the personal identity. It is a difficult task as a long as the modern cognitive science is relaying and feels relaxed on the linear algorithmic and statistical procedures. But the "rhizomality" and transgression must be incorporated in the theoretical model not as statistical noise, but as systemic factors contributing to the transformation of moral identity/moral identities of nowadays person(s) in increasingly digitized contemporary societies. It also aims to find and promote the best ways and tools for design and control those moral identities using educational strategies and tools.

In addition, the humanistic education among the scientists and engineers of the 4th Space is desperately required. The common core and ultimate outcome of the humanities are intrinsically the advancement and deepening

the self-knowledge of the felt experience of the world. Scientific and technological endeavors are positivistic in their nature and construct an alien, detached, and impersonal world. A humanistic thread is essential to entrench the technical and scientific edifice for the making the 4th Space psychologically inhabitable. The ultimate value of the humanities is the promotion of in-depth and multidimensional self-knowledge. While, as Socrates said, an unexamined life is not worth living, it is also important *how* it is examined. The scientific paradigm is unable to provide this, as long as it excludes the phenomenological approach, humanistic perspective and rhizomatous concept. There are some paths for prospective studies required for circumscribing the complex subject of moral identity development in the 4th Space and for adapting the current epistemology to this peculiar topic:

— study of rhizomality and transgressiveness of the moral identity/identities in our increasingly artifactualized/digitalised societies and their impact on the resilience and transformations of our worldviews and (moral) values: responsibility, trust, transparency, freedom, creativity;

— analysis of the possibility (opportunities and limitations) to create and secure spaces for free, critical, lucid and creative thoughts in (over)digitized societies;

— evaluation and assessment of breakthrough and major (digital) technologies and their (global and/or existential) risks in the core of societies and individual lives as risks for the future(s);

— critique of the researches, tendencies and approaches diminishing the necessity, importance and role of the humanities aggregate as well as intelligent education in critically reflecting on and actively participating in collaborative—networks and communities building;

— study of how digitalization is altering both (moral) identity and social life through the rhizomality and transgressiveness of social relations/ networks in the virtual societies;

— analysis of the effects the digitalization on the very understanding of time and space/place, as affecting our moral identity/identities;

— evaluation/assessing of digitalization affecting not just the production of data, but also its accessibility and use: the very nature of the production of knowledge and its use and sharing;

— critique of the unreflective and abusive use of AI and data mining and their effects both in research and in the education for science, moral-democratic competence, critical thinking and for creativity and innovation.

## OPEN CONCLUSIONS

The conception of this very article illustrates the pervasiveness, subtlety and utility of rhizomality as an epistemological concept for understanding the nature of the 4th Space and its semantical structure.

The moral and ethical challenges raised by technological development and the 4th Space onlife living are bigger and more complex than ever. Moral identity and morality are artifacts, products, results of our activity and cannot be anymore considered simply natural and unchangeable field(s). Such a perspective bolds asymptotically the responsibility of every human being and every society when accepting, using and promoting some of the technological changes and challenges specific for a digitized society. It also urges for the search and implementation of the best ways and methods for increasing the quality of non-formal and formal education for science, moral-democratic competence, critical thinking and for creativity at all the ages of an individual's life.

The awareness of the development of information technologies and understanding the impact they have on society will lead to a change in the concepts of the world and life in communities. Younger generations need tools for training for a constantly changing the world, with leaps in various spheres of life (economics, health, political management, education) and to train correct information skills on news in the fields of science, in order to find ways to implement them in community life. As the 4th Space and future society is continuously on the making and the human powers to foster its destiny and living environment are tremendous the risks are as high as the advantages, and will depend solely on us if we have a utopic or dystopic 4th Space for our place in the world.

## REFERENCES

American Psychological Association (n.d.), *Diffusion of Responsibility*, APA Dictionary of Psychology; https://dictionary.apa.org/diffusion-of-responsibility, accessed on February 02, 2022.

C. Baldwin, M. Greason, Caroline Hill, *Exploring the Rhizomal Self*, Narrative Works, 8 (1–2), 2018; https://doi.org/10.7202/1059846ar; accessed on 10 February 2022.

Roy F. Baumeister, *Evil: Inside Human Violence and Cruelty*, Henry Holt and Company, 2015.

M. Benedikt, *Cyberspace: Some Proposals*, in: Cyberspace: First Steps, M. Benedikt (ed.), MIT Press, Cambridge 1991, pp. 119–224.

H. Cadenas & M. Arnold-Cathalifaud, *The Autopoiesis of Social Systems and Its Criticisms*, Constructivist Foundations, 10, 2015, pp. 169–176.

J. B. Calhoun, *A "Behavioral Sink*," in: Roots of Behavior, E. L. Bliss (ed.), Harper & Brothers, New York 1962, pp. 295–315.

E. Camina & F. Güell, *The Neuroanatomical, Neurophysiological and Psychological Basis of Memory: Current Models and Their Origins*, Frontiers in Pharmacology, 8, 2017. https://doi.org/10.3389/fphar.2017.00438; accessed on 11 February 2022.

L. Cochran, *Position and the Nature of Personhood. An Approach to the Understanding of Persons*, Greenwood Press, Westport–London 1985.

G. Deleuze, F. Guattari, *A Thousand Plateaus. Capitalism and Schizophrenia,* Brian Massumi *(*trans.), University of Minnesota Press, Minneapolis–London 2005.

U. Eco, *From the Tree to the Labyrinth: Historical Studies on the Sign and Interpretation*, Anthony Oldcorn (trans.), Harvard University Press, 2014.

J. Eliot, *Models of Psychological Space. Psychometric, Developmental, and Experimental Approaches*, Springer-Verlag, 1987.
T. S. Eliot, *The Rock*, Faber & Faber, 1934.
M. Elish, *Clare Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction*, Engaging Science, Technology, and Society, 5, (2019, pp. 40–60.
D. Fasko, Jr., *Questioning and Thinking*, Inquiry: Critical *Thinking across Disciplines*, 14, 1994, pp. 43–47.
L. Floridi, J. W. Sanders, *On the Morality of Artificial Agents*, Minds and Machines, 14 (3), 2004, pp. 349–379.
L. Floridi, *A Look into the Future Impact of ICT on Our Lives*, The Information Society, 23 (1), 2007, pp. 59–64.
___ , *Infraethics–on the Conditions of Possibility of Morality*, Philosophy & Technology 30, 2017, pp. 391–394; https://doi.org/10.1007/s13347-017-0291-1; accessed on 11 February 2022.
___ , *The Informational Nature of Personal Identity*, Minds and Machines, 21(4), 2011, pp. 549–566.
___ , *Distributed Morality in an Information Society*, Science and Engineering Ethics, 19, 2012, pp. 727–743; doi: 10.1007/s11948-012-9413-4; accessed on 11 February 2022.
___ , *Commentary on the Onlife Manifesto*, in: The Onlife Manifesto: Being Human in a Hyperconnected Era, L. Floridi (ed.), Springer, 2015, pp. 21–24; https://doi.org/10.1007/978-3-319-04093-6; accessed on 11 February 2022.
M. Foucault, *The Will to Knowledge*, Penguin Books, 2006.
F. Fukuyama, *Our Posthuman Future: Consequences of the Biotechnology Revolution*, Picador, New York 2003.
S. Gálik, Sabína Gáliková Tolnaiová, *Cyberspace as a New Existential Dimension of Man*, in: Cyberspace. IntechOpen, E. Abu-Taieh, A. E. Mouatasim & I. H. Al (eds.), 2019; https://doi.org/10.5772/intechopen.88156; accessed on 11 February 2022.
C. Garnett, *Plumbing the "Behavioral Sink," Medical Historian Examines NIMH Experiments in Crowding*, Nih record, 60 (15), 2008; https://nihrecord.nih.gov/sites/recordNIH/files/pdf/ 2008/NIH-Record-2008-07-25.pdf; accessed on 11 February 2022.
F. Geyer, *Sociocybernetics*, in: International Encyclopedia of the Social & Behavioral Sciences, N. J. Smelser, P. B. Baltes (eds.), Pergamon, Oxford 2001, pp. 14549–14554.
W. Gibson, *Neuromancer*, Ace Science Fiction Special, New York 1984.
A. Giddens, *Modernity and Self-identity: Self and Society in the Late Modern Age,* Stanford University Press, Stanford 1991.
M. Grene, *Approaches to a Philosophical Biology*, Basic Books, New York 1968.
E. T. Hall, *The Hidden Dimension,* Doubleday, New York 1966.
D. Hardegger, *Towards a Merging of Spaces: A 'Holistic Concept' for the Emerging "4th Space"*, communication at IS4SI 2021. Philosophy and Computing APC; *in press* in MDPI Proceedings, 2021.
S. Harrison, P. Dourish, *Re-place-ing Space: the Roles of Place and Space in Collaborative Systems,* CSCW '96: Proceedings of the 1996 ACM conference on Computer supported cooperative work, November, 1996, pp. 67–76; https://doi.org/10.1145/240080.240193; accessed on 22 February 2022.
M. Henry, *The Essence of Manifestation*, Girard Etzkorn (trans.), Matinus Nijhoff, The Hague 1973.
D. Holmes (ed.), *Cyberspace in Glossary*, in: Virtual Politics: Identity and Community in Cyberspace, Sage, London 1997.
A. Joja, *Studii de logică* [Studies on Logic], vol. III, Editura Academiei R. S. R., Bucharest 1971.
P. Kalantzis-Cope, K. Gherab-Martín, *Emerging Digital Spaces in Contemporary Society. Properties of Technology*, Palgrave Macmillan, 2010.
I. Kant, *Critique of Pure Reason*, Paul Guyer & Allen (trans.), W. Wood Cambridge University Press, 1998.
H. Kawamoto, *L'autopoïèse et l'« individu » en train de se faire*, Revue philosophique de la France et de l'étranger, 136, 2011; pp. 347–363; https://doi.org/10.3917/rphi.113.0347; accessed on 22 February 2022.
B. Kimura, *Zur Phänomenologie der Depersonalisation*, Nervenarzt, 34 (9), 1963.
___ , *L'entre: une approche phénoménologique de la schizophrénie*, C. Vincent (trans.), Editions Jérôme Millon, 2000.

___ , *Vers une psychopathologie en première personne*, Laval théologique et philosophique, 64(2), 2008, pp. 377–385; https://doi.org/10.7202/019505ar; accessed on 22 February 2022.

___ , *Cogito and I: A Bio-logical Approach,* Philosophy, Psychiatry & Psychology, 8 (4), 2001, pp. 331–336.

K. Kircaburun, Mark D. Griffiths, *The Dark Side of Internet: Preliminary Evidence for the Associations of Dark Personality Traits with Specific Online Activities and Problematic Internet Use*, Journal of Behavioral Addictions, *7* (4), 2018, pp. 993–1003; https://doi.org/10.1556/2006.7.2018.109; accessed on 23 February 2022.

B. Latané, Steve Nida, *Ten Years of Research on Group Size and Helping*, Psychological Bulletin, 89 (2), 1981, pp. 308324; https://doi.org/10.1037/0033-2909.89.2.308; accessed on 22 February 2022.

G. Lind, *How to Teach Moral Competence*, Logos Verlag, Berlin 2019.

N. Luhmann, *The Autopoiesis of Social Systems*, in: Sociocybernetic Paradoxes—Observation, Control and Evolution of Self-steering Systems, F. Geyer, J. van der Zouwen (eds.), Sage, London 1988, pp. 172–192.

M. Madary, Thomas K. Metzinger, *Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology*, Frontiers in Robotics and AI 3, 2016; https://doi.org/10.3389/frobt.2016.00003; accessed on 20 February 2022.

H. R. Maturana, Francisco J. Varela, *Autopoiesis and Cognition: The Realization of the Living*, D. Reidel Publishing Company, Dordrecht 1980.

D. P. McAdams, *Personality, Modernity, and the Storied Self: A Contemporary Framework for Studying Persons*, Psychological Inquiry, *7* (4), 1996, pp. 295–321; https://doi.org/10.1207/s15327965pli0704_1; accessed on 22 February 2022.

A. Mihalache, *The Cyber Space-Time Continuum: Meaning and Metaphor*, The Information Society, 18, 4 (2002), 293-301, DOI: 10.1080/01972240290075138; accessed on 22 February 2022.

C. Montemayor & H. H. Haladjian, *Perception and Cognition Are Largely Independent, but Still Affect Each Other in Systematic Ways: Arguments from Evolution and the Consciousness-Attention Dissociation*, Frontiers in psychology, 8, 40 (2017). https://doi.org/10.3389/fpsyg.2017.00040; accessed on 22 February 2022.

A. Morisson, *A Typology of Places in the Knowledge Economy: Towards the Fourth Place*, in: New Metropolitan Perspectives, F. Calabrò, L. Della Spina, C. Bevilacqua (eds), ISHT 2018, Smart Innovation, Systems and Technologies, vol. 100, Springer, Cham 2019, pp. 444–451; available at SSRN: https://ssrn.com/abstract=3056754 or http://dx.doi.org/10.2139/ssrn.3056754; accessed on 22 February 2022.

D. E. Nee, Marc G. Berman, K. S. Moore, J. Jonides, *Neuroscientific Evidence about the Distinction between Short- and Long-Term Memory,* Current Directions in Psychological Science, 17 (2), 2008, pp. 102–106; https://doi.org/10.1111/j.1467-8721.2008.00557.x. accessed on 11 February 2022.

D. Norris, *Short-term Memory and Long-term Memory Are Still Different*, Psychological Bulletin, 143 (9), 2017, pp. 992–1009; https://doi.org/10.1037/bul0000108. accessed on 22 February 2022.

W. O'Donohue, K. E. Ferguson, A. E. Naugle, *The Structure of the Cognitive Revolution: An Examination from the Philosophy of Science*, The Behavior Analyst, 26, 2003**,** pp. 85–110; https://doi.org/10.1007/BF03392069; accessed on 22 February 2022.

M. J. Ostwald, *Virtual Urban Futures*, in: Virtual Politics: Identity and Community in Cyberspace, D. Holmes (ed.), Sage, London 1997.

N. J. Patel, *Reality or Fiction? Sexual Harassment in VR, The Proteus Effect and the Phenomenology of Darth Vader—and Other Stories …*, Medium, Dec 21, 2021; https://medium.com/kabuni/fiction-vs-non-fiction-98aa0098f3b0; accessed on 20 February 2022.

J. Piaget, *The Origins of Intelligence in Children,* M. Cook (trans.), W W Norton & Co., 1952; https://doi.org/10.1037/11494-000; accessed on 23 February 2022.

B. Popoveniuc, *The Psychology of the ISelf*, in: Literature, Discourses and the Power of Multicultural Dialogue, The Alpha Institute for Multicultural Studies, Arhipelag XXI Press, Tîrgu Mureş, 5, 2017, pp. 93–102.

K. R. Popper, *Epistemology without a Knowing Subject*, in: Studies in Logic and the Foundations of Mathematics, B. Van Rootselaar, J. F. Staal (eds.), Elsevier, 52, 1968, pp. 333–373.
S. Poslad, *Ubiquitous Computing. Smart Devices, Environments and Interactions*, John Wiley & Sons Ltd, 2009, pp. 17–22.
T. Postmes, R. Spears, K. Sakhel, Daphne de Groot, *Social Influence in Computer-Mediated Communication: The Effects of Anonymity on Group Behavior*, Personality and Social Psychology Bulletin, 27 (10), 2001, pp. 1243–1254; https://doi.org/10.1177/0146167201271001; accessed on 20 February 2022.
Z. R. Rowan, E. Kan, P. J. Frick, E. Cauffman, *Not (Entirely) Guilty: The Role of Co-offenders in Diffusing Responsibility for Crime,* Journal of Research in Crime and Delinquency, 2021; https://doi.org/10.1177/00224278211046256; accessed on 20 February 2022.
J. Rueda, Francisco Lara, *Virtual Reality and Empathy Enhancement: Ethical Aspects*, Frontiers in Robotics and AI, *7,* 2020; https://doi.org/10.3389/frobt.2020.506984; accessed on 20 February 2022.
G. Ryle, *Categories*, Proceedings of the Aristotelian Society, 38, 1938, pp. 189–206.
A. Sackin-Poll, *Michel Henry and Metaphysics: An Expressive Ontology*, Open Theology, 5 (1), 2019, pp. 405–419; https://doi.org/10.1515/opth-2019-0032; accessed on 20 February 2022.
J.-P. Sartre, *L'Être et le néant*, Éditions Gallimard, 1943.
D. Shin, *Empathy and Embodied Experience in Virtual Environment: To What Extent Can Virtual Reality Stimulate Empathy and Embodied Experience?*, Computers in Human Behavior, 78 2018, pp. 64–73.
J. Smith, *Geographical Rhetoric: Modes and Tropes of Appeal*, Annals of the Association of American Geographers, 86 (1), 1996.
S. L. Sorgner, *On Transhumanism*, Penn State University Press, 2020.
B. Stiegler, *Technics and Time: The Fault of Epimetheus*, Stanford University Press, Stanford 1998.
A. R. Stone, *Will the Real Body Please Stand Up?: Boundary Stories about Virtual Cultures*, in: Cyberspace: First Steps, M. Benedikt (ed.), MIT Press, Cambridge, MA 1991, pp. 94–95.
J. R. Suler, *Psychology of the Digital Age: Humans Become Electric*, Cambridge University Press, 2016.
J. Tiffin, Nobuyoshi Terashima (eds.), *Hyperreality: Paradigm for The Third Millennium*, Routledge, 2005.
Y.-F. Tuan, *Space and Place: a Humanistic Perspective*, in: Philosophy in Geography, Stephen Gale, Gunnar Olson G. (eds.), Reidel, Dordrecht 1979.
S. Turkle, *Life on the Screen*, Simon and Schuster, 2011.
S. Ventura, L. Badenes-Ribera, R. Herrero, A. Cebolla, L. Galiana, R. Baños, *Virtual Reality as a Medium to Elicit Empathy: A Meta-Analysis,* Cyberpsychology, behavior and social networking, 23(10) (2020), PP. 667–676; https://doi.org/10.1089/cyber.2019.0681; accessed on 21 December 2021.
J. Waterworth, K. Hoshi, *Human-Experiential Design of Presence in Everyday Blended Reality: Living in the Here and Now*, Springer, Cham 2016; https://doi.org/10.1007/978-3-319-30334-5; accessed on 22 February 2022.
D. M. Wegner, *Transactive Memory: A Contemporary Analysis of the Group Mind*, in: Theories of Group Behavior, B. Mullen, G. R. Goethals (eds.), Springer-Verlag, New York 1986, pp. 185–205.
M. J. Wooldridge, *An Introduction to Multiagent Systems*, 2nd ed., Wiley, Chichester 2009.

ABOUT THE AUTHOR — Professor, Ştefan cel Mare University of Suceava, Strada Universităţii 13, 720229, Suceava, Romania.
Email: bpopoveniuc@usm.ro

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Christoph M. Abels, Daniel Hardegger

# PRIVACY AND TRANSPARENCY IN THE 4th SPACE: IMPLICATIONS FOR CONSPIRACY THEORIES

doi: 10.37240/FiN.2022.10.zs.8

***ABSTRACT***

This article investigates the role of privacy and transparency in the 4th Space and outlines their implications for the development and dissemination of conspiracy theories. We argue that privacy can be exploited by individuals and organizations to spread conspiracy theories online, while organizational transparency, intended to increase accountability and ultimately trust, can have the adverse effect and nurture conspiracy beliefs. Through the lens of the 4th Space concept, we offer suggestions on how to approach those challenges which emerge as a result of the complex entanglements of both actual and virtual world across time.

**Keywords**: Transparency, privacy, disinformation, conspiracy theory, 4th space.

## 1. INTRODUCTION

With great power comes great responsibility. Although that saying holds true for many circumstances, it oversimplifies several aspects that defines the relationship between power and responsibility. Consider this question: if someone has the power to do everything he or she wants, what constrains that person in exploiting this power for solely selfish purposes? The implicit premise of the power-responsibility-relationship points to personality as a constraining factor, which works for some traits (e.g., honesty-humility), but not for others, e.g., the dark triad, consisting of Machiavellianism, Psychopathy, and Narcissism, that tends to make power exploitation more likely (Lee et al., 2013).

Aside from this, fear of social punishment might serve as a strong inhibiting factor. If powerful individuals consider other people's judgements important and fears reputational damage (or worse), they might restrain themselves from behavior that elicits contempt and subsequently social exclusion. However, if no one ever knows about certain acts or these cannot be

attributed to a specific individual, social punishment is of no concern. Beyond that, being able to conceal one's actions can even constitute a source of power, as the individual is no longer bound to normative expectations and civility, if the fear of punishment was the only thing that restrained the individual.

In the virtual world as well as the 4th Space, this power comes from the absence of attribution. When no one is able to link your online behavior to your offline identity, you are free to do whatever you want online without the fear of being punished, either legally or socially, for your misbehavior. In the actual world,[1] being able to avoid public scrutiny is seen as enabling corruption in government and corporate misconduct. However, there are reasons to obscure an individual's identity, e.g., to avoid governmental harassment in autocratic regimes. Privacy, obscuring an individual's identity, can therefore be understood as a protective layer against powerful actors allowing freedom from undue interference (Floridi, 2016). In contrast, organizations like governments and corporations are inherently more powerful than a single person. Although there are reasons for secrecy in these organizations, e.g., to allow policymakers to discuss policy solutions in private before going public, safeguarding the effectiveness of criminal investigations, or protecting trade secrets (Bok, 1989), overall organizations deserve a higher degree of scrutiny.

For governments, transparency seems inherently justified, as they are both politically accountable to their citizens as well as financially dependent on them. Helen Darbishire (2010), for example, points out that every piece of information held by public bodies should be freely accessible to all citizens, as it was created using taxpayer funds. In this context, subjecting governments to transparency-based oversight therefore seems warranted. That, however, is not sufficient. Archon Fung and David Weil (2010) argue that private sector corporations potentially also pose great risk to individuals—ranging from product safety to housing prices, mortgage rates, and "even the very soundness of the economy" (p. 109). Transparency should therefore be understood as a method to protect citizens, or help citizens to protect themselves, from these organizations.

However, in the context of conspiracy theories, both privacy and transparency can have adverse effects. As mentioned above, privacy may make people behave in a way they would normally refrain from if their behavior was on public display. Outrageous conspiracy theories might only be disseminated if the individual can be certain that it will not impact his or her offline identity, especially when the theory involves drastic accusations and includes behavior which may be strongly condemned by society. Privacy also

[1] Subsequently, we use the term "actual world" to refer to the physical space, in contrast to the virtual world that can only accessed through a device. See Wideström (2020) for a conceptual discussion.

allows foreign governments to spread propaganda or promote conspiracy theories that help their own agenda (Bok, 1989). While governments have employed these tactics at least since the early 20th century (Aaronovitch, 2010; Rid, 2020), the internet provides them with unparalleled access to citizens of other countries, and subsequently opportunities for influence operations.

For transparency, the situation is a bit more complicated. Transparency itself does not necessarily feed into the development and dissemination of conspiracy theories—how its results are used does. For governments, transparency is established through the publication of information, either proactively or via Freedom of Information (FOI) requests. Disclosure laws and regulations demand corporations to provide financial and operational information.[2] Frequently, however, whistleblowers and investigative journalists are exposing government and corporate misconduct and, eventually, induce change.[3] Given that some individuals already think that most political and corporate leaders are corrupt, working against the public interest and only concerned with their own political success, these incidents–often accompanied by large-scale media reporting and public debate, including references to past scandals–substantially feed into and bolster their mistrust. Empirical evidence indicates that people with a conspiracy predisposition, i.e., those being more likely to believe in conspiracy theories, name both "corporations and the rich" as well as, among others, governments as those "likely to work in secret against the rest of us" (Uscinski, Parent, 2014, p. 82). Additionally, as Joseph Uscinski and Joseph Parent found, people with a strong conspiracy predisposition also tend to have little trust in the government. Consequently, government and corporate misconduct fits into these people's overall image of the world: those in power will exploit us, and there is a strong body of evidence to support such claims.

The actual and virtual world reinforce each other in this respect–those who are already convinced that there are large ongoing conspiracies can go online to discuss their beliefs and further spread conspiracy theories, knowing that they do not have to fear any repercussions offline, as that their offline identity is protected through the veil of privacy. Those, however, critical about powerful organizations, might be tempted to initially engage in discussions about potential conspiracies, driven by the reporting about misconduct they encounter offline.

This complex web of interactions between actual and virtual world can be investigated through the lens of the 4th Space, which provides an analytical approach that incorporates an individual's simultaneous presence within

---

[2] The Sarbanes-Oxley Act, for example, which was enacted in 2002 after a series of major corporate and accounting scandals in the early 2000s (e.g., Enron and WorldCom), as a measure to increase the transparency of public corporations in the United States. For a summary, see Ivy Zhang (2007).

[3] For an overview of various corruption scandals, see Transparency International (2019).

a community in the actual world (e.g., a bar) and one that exists in the virtual world (e.g., Facebook), whilst also acknowledging the role of time in the interaction between these worlds (Hardegger, 2022). For example, while environments in the actual world are more likely to engage in a fact-based discussion, given the fear of social punishment for spreading falsehoods, there is little that prevents individuals from disseminating even the most outlandish claims imaginable online (Abels, 2022). In 4th Space communities, the costs of changing identities can be rather small, as individuals, e.g., can set up multiple accounts and drop those that are no longer able to gain the trust of fellow discussants.

The 4th Space understands today's information environment as a strong entanglement between an offline and online setting (actual and virtual world), along the lines of what was outlined by Zeynep Tufekci (2017), namely that "an internet society differs in significant ways from a pre-internet society, and this affects all members of that society, whether a person uses the internet or not" (p. 117). In this sense, the 4th Space provides an environment which addresses the concealment of information and its implications for the advancement of conspiracy theories in both environments.

Subsequently, we discuss the role of privacy and transparency in the 4th Space and investigate how both can nurture conspiracy theories. Starting with a brief introduction of the 4th Space, we describe the concept's elements and how it can be used to analyze interactions between the virtual and actual world. Afterwards, we contrast privacy and transparency and illustrate their role for the emergence of conspiracy theories. We conclude this paper by offering an outlook on further research questions.

## 2. THE 4th SPACE

The 4th Space is an inter- and transdisciplinary concept that provides an analytical approach for the analysis of the (emerging) hybrid society and communities.[4] It entails a methodological basis that allows to analyze and discuss how individuals and communities transcend between the actual and virtual world as well as how they interact with and among each other.

The 4th Space builds upon other concepts of community places, especially Ray Oldenburg (1989), Robert D. Putnam (2000), and Arnault Morisson (2019). Oldenburg, and subsequently Putnam, established the concept of the first, second, and third place. The first place represents an individual's home, the second its workplace. The third place, however, is the anchor of a community. It is "where you relax in public, where you encounter familiar

[4] For conceptual discussions of inter- and transdisciplinarity, see Bernard Choi and Anita Pak (2006) as well as Julie Klein (2008).

faces and make new acquaintances" (White, 2018). In his article *A typology of places in the knowledge economy: Towards the fourth place* Morrison further develops Oldenburg's concept by creating the fourth place. He argues that the knowledge economy is blurring the lines between the formerly separated places and, by doing so, establishes a fourth place that merges the other places in different configurations (Morisson, 2019).

The 4th Space concept, however, goes beyond Morisson's considerations by incorporating communities (and societies, which are constituted by the sum of communities) that emerge within and/or expand into the virtual world. These communities differ in various characteristics from (solely) actual world communities as they are not bound to a location within the actual world. Additionally, they are influenced by developments and interactions in actual and virtual world, and simultaneously influence these worlds through their individuals and/or organizations that are both present within the community. Conceptually, the 4th Space can be understood as a three-dimensional space. Each axis not only represents factors that influence and define the 4th Space but are also constitutional for the communities that emerge within this space. The three different factors are place, medium, and time.

The x or place-axis represents the most direct connection of the actual world into the virtual space and vice versa: Every individual who enters the virtual space is, in parallel, still anchored to a place in the actual world. The impressions and influences of the actual world are being taken into the virtual space and affect it through interactions of the individual with content and other individuals. Additionally, individuals' impressions of this virtual environment are also influencing the places these individuals are located in the actual world.

The medium, represented by the y-axis, is required to enter the virtual world. This includes both technical aspects, hardware and software. The medium creates heterogeneous experiences among individuals, due to software and hardware differences (e.g., accessing a community using a mobile phone in difference to a laptop or desktop computer) as well as small variations in their settings within their medium (e.g., screen brightness, adjustments of buttons, or the haptic experience that is perceived differently based on varying hand sizes of the medium's users).

The z-axis represents time. Each piece of content that is being created, amended, shared, or added, each interaction that is happening among individuals, as well as everyone that is present is doing so within a certain time (frame). Time includes the actual time as well an individual's perception of time, hence it includes the relativity of time as well. Furthermore, individuals differ in their perception on time, depending on their location as well as personal experience and expectation of time as factor.

## 3. PRIVACY & TRANSPARENCY

According to the Meriam-Webster Dictionary, something is transparent if it is "easily detected or seen through" or "characterized by visibility or accessibility of information especially concerning business practices."[5] Transparency therefore allows us to investigate the inner workings of a system or an organization. In personal affairs, this transparency is frequently avoided, as individuals want to hide certain aspects of their lives from the public (although the aspects an individual wants to hide depend on the individual itself). This privacy can be understood as "someone's right to keep their personal matters and relationships secret."[6] Frequently, transparency is discussed in the contexts of organizations, such as governments or corporations, while privacy addresses the individual level.

In the following section, we will point out how privacy and transparency are realized in the 4th Space. Starting with privacy, we illustrate how privacy differs in both virtual and actual world and how the concept of anonymity has changed online. Afterwards, we discuss the concept of transparency for corporations and governments, before relating both concepts to the 4th Space.

### 3.1. Privacy

It is an individuals' right to keep certain personal information concealed from the public and therefore prevent others from interfering in personal matters. Hence, it can also be seen as freedom from interferences and intrusion, as indicated by the Merriam-Webster definition.[7] Four kinds of freedoms can be distinguished in this respect: physical privacy, mental privacy, decisional privacy, and informational privacy (Floridi, 2016). These freedoms refer to the absence of interference or intrusion in a person's physical space, mental life, decision making, and information made accessible to the people. Luciano Floridi, however, points out that these freedoms are often intertwined, yet should be treated separately.

Lawrence Lessig (1999) takes a different approach to privacy. For him, privacy is that part of life that that is left over once everything that can be monitored (e.g., that others see or is noticed by them) or searched (all activities that create a searchable record) is subtracted. Being monitored is normal in everyday life—we are, for example, observed by other people on the streets, by security cameras, or by our neighbors. Although, our neighbors

---

[5] Merriam-Webster, *transparent*; https://www.merriam-webster.com/dictionary/transparent, accessed on 23 February 2022.

[6] Cambridge Dictionary, *privacy*; https://dictionary.cambridge.org/de/worterbuch/englisch/privacy, accessed on 23 February 2022.

[7] Merriam-Webster, *privacy*; https://www.merriam-webster.com/dictionary/privacy, accessed on 23 February 2022.

might see us at the grocery store, they rarely remember what we bought, who we talked to, or how much we eventually paid. To them, our actions are ephemeral and will not result in a lasting record.

In the face of the beginning digital transformation, Lessig argued that "we are entering an age when privacy will be fundamentally altered" (Lessig, 1999, p. 57), given that the extent to which we are monitored and information about us is becoming searchable is far greater than ever before. When shopping online, our internet provider tracks our activity, the online shop monitors what we are looking at and what we eventually bought, and the credit card company has a record of all our purchases, including the exact date and time at which we used the card. All this information is searchable and, if combined with information from other sources, might allow the creation of a personal profile that can be sold to advertising companies. Accordingly, internet users are constantly tracked online, in many cases unnoticed by the users–they become transparent for advertisers and surveillance agencies, while the people monitoring remain concealed to the users.

However, one's privacy can be protected online by masking individuals' identities, making them anonymous, using different tools like proxy servers (hiding the user's IP address behind the address of the proxy), virtual private networks (VPN, creating a secure tunnel between the server and the user's PC), as well as The Onion Router (TOR, enveloping communication between a server and the user's PC in several layers of encryption), which offers the highest level of protection (Hoang, Pishva, 2014). Additionally, privacy can be achieved through end-to-end encryption (Winkel, 2003), as used by WhatsApp and other messaging services.

The reasons for individuals to remain anonymous differ: from enabling free speech in expressive regimes (Jardine, 2018) to the creation of cryptomarkets, illicit marketplaces based on cryptocurrencies (van Hardeveld et al., 2017). Accordingly, online anonymity can be a "double-edged sword," as whilst it offers protection to whistleblowers in autocratic regimes from repercussions, but also assists individuals in avoiding criminal prosecution (Sardá et al., 2019). While being anonymous on the internet can be justified, the use of privacy-enabling technology is frequently denounced. Use of the TOR network has publicly been singled out in this respect for being associated with criminality, "characterizing it as undesirable, immoral and illegal" (Sardá, 2020, p. 257).

The use of technology to conceal an individual's identity is only one way to remain anonymous on the internet. With the emergence of Web 2.0 and the widespread adoption of social networking sites (SNS), individuals became used to create online profiles that allow them to "actively construct a representation of how they would like to be identified" (Ellison, Boyd, 2013). While some contexts demanded a clear connection between online

and offline identity, e.g., online dating (Ellison et al., 2012), others where more lenient with their identification requirements. Privacy can therefore be achieved, in some contexts, by establishing pseudonyms.

Over the course of the last years, an increasing number of popular websites have dropped anonymity and added some form of identification, some websites like Twitter, Facebook, and Instagram focused on their connection to the real-life identity through various verification systems (e.g., credit card registration). Mark Zuckerberg even promotes an idea of "radical transparency," fundamentally providing the basis for marketers to identify and predict patterns as well as to track individuals online (Kirkpatrick, 2010; Knuttila, 2011). Reportedly, Zuckerberg even told an interviewer that "having two identities for yourself is an example of a lack of integrity" (Dibbell, 2010).

However, there is (at least) one prominent website that is fully commitment to keeping their users anonymous, and therefore serves as a case for how people use this anonymity online: the imageboard 4chan. Founded in 2003 by Christopher Poole, 4chan became known to a larger audience during the 2016 US election, where its users claimed to have "actually elected a meme as president" (Ohlheiser, 2016). 4chan's anonymity is by design; accounts don't exist, only an empty name field which users do not have to fill in and if a user decides to leave it empty, 4chan assigns the account name "Anonymous" (Bernstein et al., 2011). This anonymity "makes failure cheap—nearly costless, reputation wise" (Dibbell, 2010) and allows individuals to deviate from their normal behavior, allowing them to act in ways they would never do offline, as they "can be relatively certain that their actions will not come back to haunt them" (Bernstein et al., 2011, p. 55). In a 2010 interview with the New York Times, Poole explicitly stated that he frequently received emails thanking him for providing a place in which things can be said that wouldn't be discussed with friends or family members (Bilton, 2010). According to Poole, "people deserve a place to be wrong" (Dibbell, 2010, p. 84).

Although this anonymity can serve users' freedom of expression, the lack of long-term accountability comes at a price (Knuttila, 2011). 4chan has been involved in different scandals such as Celebgate and Gamergate, both instances of sexist transgressions of the community, as well as fake bomb threats and fake trends that have contributed to a resurgence in online eating-disorder communities (Dewey, 2014). Hate speech flourishes on 4chan's /pol/ (politically incorrect) forum, which prides itself on its fight against political correctness, as racist and bigoted content surges in the aftermath of violent attacks against specific community groups, such as the 2018 Pittsburgh synagogue shooting and 2019 Christchurch mosque attacks (Malevich, Robertson, 2020; Thompson, 2018; Zelenkauskaite et al., 2020).

4chan is furthermore seen by some scholars as the origin of conspiracy theories such as "Pizzagate" (Tuters et al., 2018). Pizzagate describes a con-

spiracy theory based on private e-mails belonging to Hillary Clinton's former campaign manager John Podesta, which were leaked by Wikileaks during the campaign phase of the 2016 US Presidential election. Users on /pol/ manufactured bogus claims about Podesta and other high-profile members of the Democratic Party being involved in a satanic pedophilia ring operated out of a Washington D.C. pizzeria. The conspiracy theory subsequently spread beyond 4chan to, among others, Facebook and Twitter as well as Turkish pro-government media outlets (Tuters et al., 2018; Wendling, 2016). This transition from 4chan to other SNS is no isolated incident (Zannettou et al., 2017). More recently, Pizzagate has re-emerged on TikTok, a SNS focused on short videos that has become popular since its 2016 launch and has a large global network of members, now including a variety of business, political and cultural elites, such as Bill Gates, Oprah Winfrey, and Ellen DeGeneres (Kang, Frenkel, 2020).

These examples illustrate the dark sides of privacy. While individuals have various legitimate reasons to protect their privacy online, e.g., to express their beliefs without fear of political oppression in autocratic regimes, some individuals use privacy, through means of anonymity, to disseminate hate speech and amplify conspiracy theories. This problem is not confined to 4chan with its focus on anonymity. Other SNSs such as Facebook, Instagram, Twitter and YouTube also struggle to contain the spread of disinformation. A number of SNSs enable users to register under a pseudonym, thus they are able to avoid being identified and held liable for their online behavior. And even if the account is banned, individuals are able to circumvent the ban by creating new accounts, although thereby violating Twitter's use policy (Twitter, 2020). The lack of accountability, that is associated with anonymously or pseudonymously posting content online, can therefore be seen as a reason for the surge of disinformation online.

### 3.2. Transparency

As previously stated, whilst privacy refers to the individual level, transparency is frequently discussed at the organizational level. In this respect, both concepts differ in their respective goals: with organizations frequently being more powerful than individuals, privacy protects individuals from those organizations, but also from the interference of other individuals, while transparency sheds light on the potential wrongdoings of these organizations. Accordingly, transparency can be seen as "the ability to look clearly through the windows of an institution" (den Boer, 1998, p. 105) while Albert Meijer (2009, p. 258) phrases it as "the general idea that something is happening behind curtains and once these curtains are removed, everything is out in the open and can be scrutinized". We will subsequently focus on two types of organizations for which transparency plays an important role, governments and corporations.

At the heart of the issue lies an inherent information asymmetry: government officials and corporate executives have direct access and control over the actions of their respective organizations, their assets and funds and other resources. From the perspective of principal agent theory, the principal (citizens, shareholders or stakeholders) delegate certain tasks to an agent who's interests either align with those of the principal, but frequently deviate from them (Jensen, Meckling, 1976). In these cases, the agent can exploit this information asymmetry for its own gains. This asymmetry holds for both relationships between citizens and the government as well as between corporate executives and shareholders (Stiglitz, 2002). This frequently results in corporate misconduct (Heath, 2009) and corruption, understood here as the misappropriation of state resources for the private gains of politicians and bureaucrats (Mungiu-Pippidi, 2006; 2013). Already in 1914, Louis Brandeis stated that "sunlight is said to be the best disinfectants; electric light the most efficient policemen" (Brandeis, 1914), making the case for transparency as an effective tool against those acting outside of the public eye. Establishing this transparency, however, comes at a cost, as measures need to be put in place to enable the principle to collect the necessary information that eventually makes the organization more transparent.

For corporations, transparency touches upon a variety of different areas– from financial disclosure to product safety requirements (Fung et al., 2007; Hermalin & Weisbach, 2007). Different approaches to corporate governance as measures to enable the principal to better control the agent have been developed (for an overview, see Anheier, Abels, 2020). Additionally, regulatory bodies and watchdog organizations take interest in corporate behaviour and data published by the corporations, adding an additional layer of corporate oversight. Still, in recent years there have seen a series of scandals that have revitalised interest in organisational transparency, e.g., the bankruptcy of financial service provider Wirecard (Barnert, 2021), various scandals related to privacy and mental health at Facebook (Vaidhyanathan, 2018), including the infamous Cambridge Analytica scandal (Granville, 2018), the crash of two Boeing 737 MAX that killed 346 people (Robinson, 2021) as well as the defrauding of customers and investors by the now defunct biotechnology company Theranos (Carreyrou, 2018). These examples illustrate the limits of transparency for corporate control, as frequently regulatory bodies fail to act upon their mandates and investigate problems, often with a human cost. In some cases, such as Wirecard, it was investigative journalists who disclosed the company's misconduct and prompted a broader investigation by German authorities (Storbeck, 2021).

While corporate scandals harm customers and shareholders, corruption of government officials can, aside from the immediate financial harm, damage citizens' trust in their leaders. In governments, transparency therefore serves as a constraint against corruption (Mungiu-Pippidi, 2015), but also as a way

to assess government performance (Stiglitz, 2002). Transparency is therefore often discussed in the context of government accountability. Yet, beyond that, access to information held by government or government agencies is increasingly seen as a human right, as an increasing number of constitutions and international courts have enshrined this right into treaties related to freedom of expression and information provision (Darbishire, 2010). To illustrate this development: at the time the Berlin Wall fell in 1989 only 12 countries had "access to information" or "Freedom of Information" laws (FOI), primarily in states with longer-established democracies (Darbishire, 2010). By 2019, 119 nations had implemented FOI laws (Feldman, 2019).

There are two ways on how the public can access information held by public institutions. Citizens can either submit requests for information (reactive disclosure) under FOI laws, or access via those institutions which proactively publish information without such requests. The result of this proactive disclosure is proactive transparency, which makes it more complicated for officials to manipulate information. Proactive transparency is especially effective in authoritarian regimes, where citizens lacking the necessary power to protect themselves from government misconduct or worse, might otherwise unable to request information which might expose vested interests of certain actors (Darbishire, 2010).

However, processes such FOI do not necessarily produce more actual transparency. As Hood (2007) points out, if politics and bureaucracy show a certain orientation for blame-avoidance, behavioral patterns can be observed that create circumstances in which different strategies are employed to limit the blame actors can receive, frequently with negative consequences for transparency. In the face of intentional maneuvering of bureaucrats and politicians to avoid blame by, among other approaches, reducing the degree of transparency through means such as unintelligible records of meetings (e.g., in form of PowerPoint presentations), telephone calls, or in person discussions that are not recorded at all (Hood, 2007), citizens who expect the state (elected politicians and bureaucrats) to work to increase their quality of life might end up frustrated and lose trust in their leadership. Furthermore, if bureaucrats choose to sabotage initiatives that would increase transparency and thereby accountability, it is unsurprising if citizens want to know what accountability these bureaucrats seek to avoid–and subsequently assume the worst.

### 3.3. Privacy and Transparency in the 4th Space

Privacy and transparency have several implications for the 4th Space. The means through which individuals enter the virtual world, the medium, remain largely opaque to many users—they lack, for example, the technical expertise to fully understand the device they use, how the software works,

who might be able to eavesdrop, and what data is collected. Accordingly, their degree of privacy differs substantially dependent on their understanding of the respective technology, which is in many cases superficial at best (Park, 2013). As a result, individuals might expect their privacy to be more strongly protected than it *de facto* is. In respect to anonymity, as a measure to establish privacy, expectations about the absence or presence of anonymity or pseudonymity might also differ from the situation individuals encounter in the communities they engage in. While some communities have implemented means to verify one's true identity, e.g., credit card registrations, others lack these approaches, despite existing platform policies that suggest the need for identity verification processes. Figure 1 illustrates these deviations.

| Anonymity allowed by platform provider | Anonymity expected by platform user: Yes | Anonymity expected by platform user: No |
|---|---|---|
| Yes | 4chan<br>Discord<br>Reddit<br>Twitter | LinkedIn |
| No | Facebook | Tinder |

**Figure 1.** Differences in expected and allowed privacy in online communities

On Facebook, for example, given its policy that demand registration with one's true name (Facebook, n.d.), individuals might expect every account on the platform to be connected to a similar identity in the actual world. However, there is little enforcement of the policy by Facebook, allowing individuals to use pseudonyms without disclosing this to other individuals. As a result, Facebook users must maintain a certain situational awareness when engaging with others on the platform, given that the lack of long-term accountability associated with pseudonymity can increase the chance of encountering individuals with potentially malign intentions, e.g., to spread conspiracy theories and spreading disinformation or engaging in cybercrime. On Twitter, however, as the self-ascribed "free speech wing of the free speech party" (Halliday, 2012), individuals cannot expect fellow users to disclose their actual world identity. Still, Twitter, among other SNS, has introduced a verification check for "accounts of public interest," e.g., journalists, government officials, and prominent persons, to increase the trust between users (Twitter, n.d.). However, this does not address the issue of

manipulation, as this only helps to identify those individuals which may want to be identified. Online dating sites, such as Tinder, have struggled for a long time with fake user profiles and being used for online scams (Drouin et al., 2016; Murphy, 2016).

Time and engagement might furthermore implicitly increase the vulnerability towards manipulation attempts: the more information about a person becomes public, due to a lack of privacy or misunderstandings about the identity of other members in a community, the easier it is for malicious actors to exploit this information for manipulative or harmful purposes. This has been, for example, seen in recruitment efforts for ISIS (Callimachi, 2015), but also fraudulent online dating scams (Whitty, Buchanan, 2016). While time might lead to an unnoticed accumulation of information that allow for an identification of a person, this information can also be intentionally made public by others. This so called "doxing", understood as the unvoluntary disclosure of private information by a third party, can concern information related to a person's identity, location, or supposedly immoral activity (Douglas, 2016). Depending on its purpose, doxing can be "a tool for establishing accountability for wrongdoing, a means of intimidation and incitement to cause harm, and a way of silencing minority or dissenting views" (Douglas, 2016, p. 209). From the perspective of the 4th Space, doxing bridges the gap between the actual and the virtual world (involving all three axes, medium, place, and time), thereby negating any individuals' attempts for privacy protection.

Additionally, an individual's location (the place axis) can impact the relevance of understanding and protecting their privacy. Depending on political (autocracy vs. democracy) and cultural (liberal vs. conservative) contexts, some opinions can only be safely expressed or information obtained when one's true identity remains unknown. Correspondingly, in oppressive autocratic regimes, privacy might be a question of mitigating the potential for physical harm when speaking out against the government, yet sometimes virtual environments may simply offer a space for discussions on topics that are otherwise off-limits due to cultural sensitivities wherever an individual is situated, e.g., discussing marital issues or homosexuality in China (Wang, 2013). These social norms are highly context dependent, sometimes with differences within a country (e.g., abortion in rural and urban Germany) or between neighboring countries (e.g., assisted suicide in the Netherlands or France) and illustrate how physical location impacts the role of virtual communities that allow a certain degree of privacy for the free expression of thoughts (Tufekci, 2017). As a result of individuals being present in both the virtual and the actual world, transparency of governments and corporations also impact their perception of the actual world, which can spillover into the virtual environment.

## 4. CONSPIRACY THEORIES

Privacy and transparency can be seen as two sides of one coin: Privacy allowing individuals to act in concealment, transparency lifting that veil behind organizations could hide their actions. Both concepts therefore are related to the idea of secrecy–an important element of many definitions of conspiracy theories. Accordingly, Cass Sunstein and Adrian Vermeule (2008) define them as "an effort to explain some event or practice by reference to the machinations of powerful people, who have also managed to conceal their role" (p. 4). Brian L. Keeley (1999) focusses on the agents causing the event in question, seeing a conspiracy theory as "a proposed explanation of some historical event (or events) in terms of the significant causal agency of a relatively small group of persons–the conspirators–acting in secret" (p. 116).

However, definitions of conspiracy theories can also highlight their explanatory role. They can be understood as any explanation of an event that invokes a conspiracy as its cause (Dentith, Orr, 2018) or, in the sense of David Aaronovitch (2010), as "the attribution of deliberate agency to something that is more likely to be accidental or unintended" (p. 6). He furthermore expands his definition by arguing that the secret actions of the persons identified by the conspiracy theory as perpetrators are more reasonably explained by those that had overtly acted.

The conspirators themselves are also subject of debate. According to Keeley (1999), the group causing the event does not need to be powerful, its pivotal role is sufficient. But, given the limited power of conspirators, secrecy is needed to execute the conspiracy: If they would act in public, their plans would be obstructed by others. In the face of the pivotal role many conspiracy theories assume, a single person is frequently not enough for a conspiracy. Hence, as Matthew Dentith and Martin Orr (2018, p. 441) point out, a conspiracy is a social relationship–although fragile one, in which there is "a potential leaker, a potential whistleblower, and a potential turncoat."

In the following section, we discuss how transparency can advance conspiracy theories. Afterwards, we highlight the role of the 4th Space as an analytical framework to investigate these theories.

### 4.1. The role of transparency in advancing conspiracy theories

Communities in which conspiracy theories flourish are paradoxical: While some of the actors make outrageous claims that probably do not seem believable to most persons, fellow conspiracy theorists in many cases do not seem overly concerned with the validity of these claims. Although these people do

not trust the government and do not trust those attempting to debunk conspiratorial claims with evidence, they still believe in what are frequently unwarranted theories about covert agents acting against the common good. If their distrust is high enough to consider almost everything to be a conspiracy theory, except those theories most worthy of that description, why do they trust an anonymous person in an opaque virtual environment?

Government transparency may be one of the reasons. As several authors have pointed out (Fung, 2013; Margetts, 2011), making governmental action transparent, especially for the sake of accountability, might lead citizens to focus on missteps, policy failures, and corruption, in a manner that Fung and Weil (2010, p. 106) call "gotcha game." Citizens are actively looking for failures of those in power and feel confirmed once they found something. Some of the conspiracy communities, truthers, even refer to themselves explicitly as those that look behind the curtain and expose what is concealed by the government (Kay, 2011). As Bok (1989) has pointed out, secrecy–understood as the result of concealment—is strongly linked in many people's minds with deception. Beyond that, the idea frequently prevails that secrecy itself is discreditable, as people only conceal what they find "shameful or undesirable" (p. 8). Trying to expose the secrets governments hide from the people is therefore a value in itself, as only that is concealed which is diametral to the greater good of society.

With an ever-growing number of leaks from whistleblowers–Panama Papers, Pandora Papers, Paradise Papers, and very recently Suisse Secrets–conspiracy theorists can easily feel vindicated in their believe about widespread corruption of those in power. As evidenced by the recent Suisse Secret leaks, that shed light on one of the world's most important financial institutions, the Credit Suisse bank, there indeed is a powerful elite, ranging from the son of an Azerbaijani strongman, Egyptian intelligence officials, to various wealthy criminals, that is protected by laws and catered by institutions that both support and profit from them (OCCRP et al., 2022). Decades earlier, Tobacco corporations were either deliberately concealing or at least whitewashing the negative health consequences of smoking (Rabin-Havt & Media Matters, 2016). These are only some examples that illustrate the large number of scandals involving governments or corporations. Jointly, these incidents can undermine public trust and lead people in the hands of those sharing and subsequently nurturing their mistrust.

As a result, although transparency should make governments more accountable to the public, and through this accountability increase trust in their doings, the opposite can be the case, if transparency focusses public attention on misconduct by government officials. In the private sector, transparency can uncover criminal or norm-violating behavior, e.g., environmental pollution, corruption, customer endangering. In combination, both mechanisms can reinforce individuals' perception of powerful actors,

being it government or corporate leaders, frequently engaging in behavior that is in direct opposition to the public good.

### 4.2. The 4th Space as analytical framework for virtual conspiracy communities

Conspiracy theories seem to accompany humanity through its history. As Uscinski and Parent (2014, 3) point out, "naturally, conspiracy theories flourish across space just as much as they do across time." Yet, while they seem to be ever-present–from the antisemitic *Protocols of the Elders of Zion* in the 20th century (Aaronovitch, 2010) to the recent conspiracy theories involving the Covid-19 pandemic (Uscinski et al., 2020)–times of anxiety, paranoia, and a perceived loss of control in large parts of society seem to be the ideal environments for conspiracy theories (Douglas et al., 2019; van Prooijen, Douglas, 2017).

Hence, conspiracy theories largely appear to be a response to a state of crisis. However, today's information environment allows actors to spread conspiracy theories for other purposes, e.g., because enjoy doing it (Buckels et al., 2014) or engage in state-sponsored disinformation operations (Rid, 2020). A substantial driver of this are virtual communities, in which individuals can encounter conspiracy theories and discuss them with like-minded others.

The 4th Space offers an analytical framework to investigate these encounters. Starting with the medium, individuals have a great choice of virtual communities they can engage with to exchange views on conspiracy theories and encounter new ones. From Facebook to Twitter and TikTok, in simple terms, every SNS proffers content that promotes conspiracy beliefs. Individuals can therefore not only engage with one community on a single platform but can be part of several discussions across platforms, thereby exchanging content between platforms. This is frequently seen on WhatsApp and Telegram, where links to YouTube and other Websites are shared. Given that not every individual is present on the majority of SNS, through this interconnection of virtual communities these individuals are still likely to encounter the most prominent conspiracy theories.

In the 4th Space, individuals are however not only exposed to information from the virtual world, as the remain anchored through their location in the physical world. Given conspiracy theories are oftentimes explanations for significant historical or political events, individuals are likely to discuss those events with their immediate social environment—at work, home, or in bars and restaurants, talking with their friends and family. Yet, as these individuals can be simultaneously present in their virtual communities, discussions from the actual world can migrate to the virtual one and vice versa. Assuming that the belief in conspiracy theories is by many seen as a deviation from normal behaviour—some authors even view conspiracy

belief as pathological (Hofstadter, 1965)—individuals might refrain themselves from disclosing their true beliefs about certain events, due to the fear of being socially stigmatized or excluded. The privacy of the virtual space can provide the necessary safety to freely articulate their views.

Location can also have a direct impact on the information entering the virtual space. Proximity to sites of emergencies can increase the quality of information shared online (Starbird, Palen, 2010; Thomson et al., 2012). The opposite was observed in New York City after the attacks on the World Trade Center: residents in the city strongly believed that the government knew about the attack in advance and failed to act, while this belief was less prevalent in the rest of the US (Sunstein, Vermeule, 2008).

If these individuals now encounter a government or corporate scandal, they can discuss the matter with their immediate environment in the virtual world, maybe also just learn about them from friends and family, and carry it over to their virtual communities. In these communities, they can then elaborate on the underlying causes of the scandal and investigate what the media, which frequently uncovers these scandals, has (deliberately) left unreported. Yet, their beliefs might remain concealed to both their actual social environment and the virtual one.

The 4th Space's third component, time, underlines the role of technological progress, the durability of conspiracy theories, and their long-term impact. As Uscinski and Parent (2014) point out, some individuals might be socialized into a worldview that has a stronger emphasis on conspiratorial thinking. One driver of that is today's presence of the high-choice media environment. While it was difficult in the past to encounter media that caters to certain ideologies and reinforces them, individuals can nowadays choose the media outlet that suits their ideological preferences best (Van Aelst et al., 2017). Additionally, with an increasing lifespan, individuals are also potentially more likely to experience a conspiracy.

Virtual environments furthermore make discussions less ephemeral. Conversations at work or in a bar do normally not leave a record.[8] On Facebook or Twitter, for example, every discussion and exchange with other users create a searchable record, until the users decide to delete it. However, even then, other users might have made a screenshot of the conversation and uploaded it to a Cloud server. Even in communities that are deliberately designed to be ephemeral, like 4chan, users can easily create copies of that conversation and therefore expand its lifespan.

[8] Sometimes, however, discussions might be recorded, intentionally or unintentionally. Yet, this is not what most people would expect nor how most situations are set up.

## 5. DISCUSSION

These examples show the complex web of interconnections between place, medium, and time that constitutes the 4th Space. They furthermore indicate how the 4th Space can be used as an analytical framework to investigate the implications of these interdependencies for the development and dissemination of conspiracy theories.

Beyond that, the 4th Space can be used to identify solutions to cope with the problems identified in this article. As already mentioned, there is little understanding of technology—the medium—on the part of the users who move, exchange, and create or use content in 4th Space. If, for example, users acquire a new hardware or software currently in use receives an update, they might be unaware of the impact on their privacy. This is understandable, as the speed of technological changes, especially in software, might overwhelm most users and pose an enormous task even for the more experienced ones. Nevertheless, it must be emphasized that a fundamentally better understanding of the necessary technology in the 4th Space and the impact on the user's experience within the 4th Space, especially regarding privacy, would also likely create more trust in the interaction.

The medium's affordances play a crucial role in this respect. For instance, as users can have multiple accounts in a certain 4th Space, the general idea of "one body, one identity" (Donath, 2020) does not apply in that specific environment. If these users are not aware of this or policies exist that create the illusion that every online persona is connected to a similar offline one, although the policy is not enforced (e.g., as it is frequently the situation on Facebook), users might be misled in their interaction with other individuals about their true identities. This makes it more difficult to identify incorrect information or even targeted false messages, as users' experiences on online dating sites make clear (Rege, 2009). To counteract this, virtual communities should incorporate design features that make the state of privacy policies more salient and support the situational awareness of their members. In the context of Covid-19-related disinformation, several SNS have added labels and other warning mechanisms to their platforms in order to protect individuals from falling for misleading information (Bond, 2020). A similar approach could be taken to increase an organization's transparency in relation to when and how they enforce existing privacy policies.

In the context of conspiracy theories, the role of anonymity, as a tool to establish privacy, needs to be discussed as well. Every 4th Space can be divided into one of three categories: full anonymity, partial anonymity, and no anonymity. The former category includes 4chan, where no registration of any kind is expected and the users themselves respect the anonymity and thus the privacy of others. On the contrary, any kind of connection to the actual world would ultimately undermine the basic idea of the 4th Space

that 4chan creates. At the same time, this also means that a user must take any information and interaction within the respective community with a grain of salt. The second category, partial anonymity, includes 4th Spaces such as reddit or Twitter. These require a registration for the interaction in them and thus also the deposit of corresponding data, such as a mail address. But there is no obligation to verify the actual world identity, and every user can create and use an unlimited number of accounts. Just as with 4chan, every user of these 4th Spaces must assume that here, too, every piece of information does not necessarily have to correspond to the facts. The last group, those that do not guarantee anonymity and want to combine the actual world identity with the virtual world identity, includes Facebook and LinkedIn. While the latter merely carries this claim with it, at least Facebook is also trying to enforce it legally, albeit not successfully. In the case of Facebook, this leads to the paradoxical situation that the platform's affordances tend to signal users to assume that most information on Facebook comes from real people and organizations, thereby creating the impression of accountability for the spread of misleading information, but at the same time claim to be allowed to remain anonymous. However, given that reality distortion might be the norm online, the mere perception of a user being who he or she claims to be is not enough to take the validity of information for granted (Zimbler, Feldman, 2011).

On the place axis, the situation is more complex. Privacy and transparency are not merely technical matters, but subject to legal, cultural, linguistic, and other factors. For instance, whether privacy is perceived as valuable depends on the individual's location. Privacy is certainly more helpful in oppressive autocracies, in which exercising free speech might pose an immediate threat to individuals well-being. Transparency of government and corporations also differs across countries, as some nations, although enacting FOI laws, have little interest in becoming more transparent. This difference between *de jure* and *de facto* transparency has been the cause for the development of new indicators to assess a government's objective level of transparency (Mungiu-Pippidi, Dadašov, 2016). Beyond that, whether the information provided by a government can be used to hold it accountable depends on the existence of civil society actors capable of analyzing the data and advocating for change (Fung, 2013). The same is true for corporations, as misconduct is frequently exposed by investigative journalists, e.g., in the case of Theranos, which has defrauded customers as well as investors (Carreyrou, 2018).

However, the push towards good governance through transparency can be act as a double-edged sword: although transparency can achieve greater trust in government and bureaucracy, repeated exposure to strategies to undermine these initiatives can have a lasting negative impact on trust in government, potentially increasing the likelihood of citizens to adopt conspiracy theories. In combination with large-scale leaks from whistleblowers

that expose the wrongdoings of powerful elites, a generalized mistrust towards anyone in power can be the result, providing fertile ground for conspiracy theories to flourish.

At the same time, also in contrast to the medium axis, individuals in some locations tend to have greater awareness of the importance of privacy and transparency, since the connection to the actual world of the respective users is much more direct here. Accordingly, there is a more reflection on the role of privacy and transparency in the 4th Space. Nothing illustrates this better than the debate surrounding the General Data Protection Regulation (GDPR), which was ultimately not a technical discussion, but a transfer of the European self-image of privacy and transparency into the 4th Space (Greenleaf, 2012). This underlines that the 4th Space is not limited to a single, clearly defined geographic area, but encompasses every place where users log in. Accordingly, different legal concepts of privacy and transparency from the actual world, but also socio-cultural, linguistic, economic, and religious ones compete in the 4th Space, each depending on the individual background of the users and the location in which they are located.

For example, a 4th Space may be primarily used by users and hosted by an organization from North America and Europe. However, as soon as a user from a completely different region, such as South Africa, enters and becomes part of the 4th Space, their respective definitions of privacy and transparency also become part of it, thus expanding the 4th Space on the place axis accordingly. At the same time, however, this user is also influenced by the already existing definitions of privacy and transparency within the 4th Space, which again impacts the individuals' actual world environment.

Which brings us back to the medium axis. Because even if legal and societal changes in the understanding of privacy and transparency in the 4th Space are possible, corresponding adjustments and improvements are also necessary on the technical and design level. It would, e.g., make sense on the part of those who are technical responsible for the respective 4th Spaces to create the possibility of more clearly tracing the development of information and discourse within the community. 4th Spaces such as Reddit and Twitter are less prone to be undermined by conspiracy theories, since here, a) the history of discourses can be tracked more directly, b) the community itself actively evaluates and shares information, and c) individuals know that there is no requirement of connecting the actual world to the virtual world identity, so they usually take every information with the required skepticism (Cinelli et al., 2021; Theocharis et al., 2021).

Our remarks here pose several questions that deserve further investigation. Concerning the medium, the issue of privacy might evolve over the course of the next years, given technological developments around deepfakes, manipulated multimedia content (Chesney, Citron, 2018; Verdoliva, 2020), and Facebook's so-called Metaverse. Through deepfakes, individuals

can, for example, alter videos about themselves to conceal their identity from others while pretending to show their true self. In the Metaverse, the increased degree of interaction, including virtual avatars that represents a person's behavior more directly, might alter privacy, as it is easier to observe patterns of behavior, speech, and other aspects that are more difficult to obscure. Accordingly, how these thinner privacy affects the spread of conspiracy theories in the Metaverse should be subject to future research.

Beyond that, further research is needed to investigate how conspiracy theories move from the actual world to virtual communities and vice versa. Although it is arguably more likely that individuals discuss these issues online, there is an increasing number of examples in which virtual communities around disinformation reach over into the actual world–the Querdenker movement in Germany, which largely organizes itself via Telegram and other platforms, is only one of the more recent examples (Koos, 2021). Other instances include Pizzagate (Tuters et al., 2018) and the storm on the US Capitol on January 6, 2021 (Barry et al., 2021).

Finally, the question of how government transparency can lead to the emergence or advancement of conspiracy theories demands further attention. While it seems intuitively convincing that the gotcha game (Fung, Weil, 2010) can lead to the emergence of conspiracy beliefs, as information are interpreted in the face of pre-existing beliefs and attitudes (Miller et al., 2016; Taber et al., 2009) and subsequently twisted and turned to fit into conspiracy beliefs (which is what happened in the case of Pizzagate on 4chan), the literature has so far hardly addressed this issue.

The 4th Space provides a holistic framework to analyze and combat the spread and development of conspiracy theories, by incorporating aspects of place, medium, and time. Following the Swiss Cheese model to mitigate disinformation online (Bode, Vraga, 2021), the 4th Space framework can help us to identify the relevant protective layers and allow us to shed light on the man or woman behind the curtain.

## REFERENCES

D. Aaronovitch, *Voodoo Histories: The Role of the Conspiracy Theory in Shaping Modern History*, Riverhead Books, New York 2010.

C. M. Abels, *Everybody Lies: Misinformation and Its Implications for the 4th Space*, Proceedings, 68, 2022, in press.

H. K. Anheier, C.M. Abels, *Corporate Governance: What Are the Issues?*, in: Advances in Corporate Governance: Comparative Perspectives, H. K. Anheier, T. Baums (eds.), Oxford University Press, Oxford 2020, pp. 10–42.

J.-P. Barnert, *Wirecard Chapter Ends With Stock Set to Delist From Exchanges*, Bloomberg, 2021; https://www.bloomberg.com/news/articles/2021-11-12/wirecard-chapter-ends-with-stock-set-to-delist-from-exchanges; accessed 03 March 2022.

D. Barry, M. McIntire, M. Rosenberg, *"Our President Wants Us Here": The Mob That Stormed the Capitol*, The New York Times; https://www.nytimes.com/2021/01/09/us/capitol-rioters.html; accessed 03 March 2022.

M. Bernstein, A. Monroy-Hernández, D. Harry, P. André, K. Panovich, G. Vargas, *4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community*, Proceedings of the International AAAI Conference on Web and Social Media, 5 (1), 2011, pp. 50–57.

N. Bilton, *One on One: Christopher Poole, Founder of 4chan*, The New York Times, 2010, https://bits.blogs.nytimes.com/2010/03/19/one-on-one-christopher-poole-founder-of-4chan/?_php=true&_type=blogs&_r=0; accessed on 03 March 2022.

L. Bode, E. Vraga, *The Swiss cheese model for mitigating online misinformation*, Bulletin of the Atomic Scientists, *77*(3), 2021, 129–133. https://doi.org/10.1080/00963402.2021.1912170

S. Bok, *Secrecy: On the Ethics of Concealment and Revelation*, Vintage Books, New York, 1989.

S. Bond, *Twitter Expands Warning Labels To Slow Spread of Election Misinformation*, NPR, 2020; https://www.npr.org/2020/10/09/922028482/twitter-expands-warning-labels-to-slow-spread-of-election-misinformation?t=1648567293523; accessed on 04 March 2022.

L. Brandeis, *Other People's Money and How the Bankers Use It*. Frederick A. Stokes, New York, 1914.

E. E. Buckels, P. D. Trapnell, D. L. Paulhus, *Trolls just want to have fun*, Personality and Individual Differences, *67*, 2014, 97–102; https://doi.org/10.1016/j.paid.2014.01.016

R. Callimachi, *ISIS and the Lonely Young American*, New York Times, 2015; https://www.nytimes.com/2015/06/28/world/americas/isis-online-recruiting-american.html?searchResultPosition=1

J. Carreyrou, *Bad Blood: Secrets and Lies in a Silicon Valley Startup*. Penguin Random House, New York 2018.

R. Chesney, D.K. Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, No. 692; Public Law Research Paper), 2018; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954

B. C. K. Choi, A. W. P. Pak, *Multidisciplinarity, Interdisciplinarity and Transdisciplinarity in Health Research, Services, Education and Policy: 1. Definitions, Objectives, and Evidence of Effectiveness*, Clinical and Investigative Medicine, 29 (6), 2006, pp. 351–364; http://www.ncbi.nlm.nih.gov/pubmed/17330451

M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, M. Starnini, *The Echo Chamber Effect on Social Media*, Proceedings of the National Academy of Sciences, 118 (9), 2021; https://doi.org/10.1073/pnas.2023301118

H. Darbishire, *Proactive Transparency: The Future of the Right to Information?*, The World Bank, Washington, DC 2010; https://openknowledge.worldbank.org/bitstream/handle/10986/25031/565980WP0Box351roactiveTransparency.pdf?sequence=1&isAllowed=y; accessed 21 February 2022.

M. den Boer, *Steamy Windows: Transparency and Openness in Justice and Home Affairs*, in: Openness and Transparency in the European Union, V. Deckmyn, I. Thomson (eds.), European Institute of Public Administration, Maastricht, 1998, pp. 91–105.

M. R. X. Dentith, M. Orr, *Secrecy and Conspiracy*, Episteme, 15 (4), 2018, pp. 433–450; https://doi.org/10.1017/epi.2017.9

C. Dewey, *Absolutely Everything You Need to Know to Understand 4chan, the Internet's Own Bogeyman*, The Washington Post, 2014; https://www.washingtonpost.com/news/the-intersect/wp/2014/09/25/absolutely-everything-you-need-to-know-to-understand-4chan-the-internets-own-bogeyman/; accessed 18 February 2022.

J. Dibbell, *Radical Opacity*, Technology Review, 113 (5), 2010, pp. 82–86.

J. S. Donath, *Identity and Deception in the Virtual Community*, in: Communities in Cyberspace, P. Kollock, M. Smith (eds.), Routledge, London 2020, pp. 37–68; https://doi.org/10.4324/9780203194959-11

D. M. Douglas, *Doxing: a Conceptual Analysis*, Ethics and Information Technology, 18 (3), 2016; 199–210. https://doi.org/10.1007/s10676-016-9406-0

K. M. Douglas, J. E. Uscinski, R. M. Sutton, A. Cichocka, T. Nefes, C.S. Ang, F. Deravi, *Understanding Conspiracy Theories*, Political Psychology, 40 (S1), 2019, pp. 3–35; https://doi.org/10.1111/pops.12568

M. Drouin, D. Miller, S. M. J. Wehle, E. Hernandez, *Why Do People Lie Online? "Because Everyone Lies on the Internet,"* Computers in Human Behavior, 64, 2016, pp. 134–142; https://doi.org/10.1016/j.chb.2016.06.052

N. B. Ellison, D.M. Boyd, *Sociality Through Social Network Sites*, in: W. H. Dutton (ed.), Sociality Through Social Network Sites, Oxford University Press, Oxford, 2013. https://doi.org/10.1093/oxfordhb/9780199589074.013.0008

N. B. Ellison, J. T. Hancock, C. L. Toma, *Profile as Promise: A Framework for Conceptualizing Veracity in Online Dating Self-presentations*, New Media and Society, 14 (1), 2012, pp. 45–62; https://doi.org/10.1177/1461444811410395

Facebook, *What Names Are Allowed on Facebook?,* n.d.; https://www.facebook.com/help/112146705538576; accessed 07 March 2022.

S. Feldman, *Where Do Freedom of Information Laws Exist?* Statista—The Statistics Portal, 2019; https://www.statista.com/chart/17879/global-freedom-of-information-laws/; accessed 20 February 2022.

L. Floridi, *The 4th Revolution: How the Infosphere is Reshaping Humanity*, Oxford University Press, Oxford 2016.

A. Fung, *Infotopia: Unleashing the Democratic Power of Transparency*, Politics and Society, 41 (2), 2013, pp. 183–212; https://doi.org/10.1177/0032329213483107

A. Fung, M. Graham, D. Weil, *Full Disclosure: The Perils and Promise of Transparency*, Cambridge University Press, Cambridge 2007.

A. Fung, D. Weil, *Open Government and Open Society*, in: Open Government, D. Lathrop, L. Ruma (eds.),O'Reilly Media, Sebastopol, 2010, pp. 105–114.

K. Granville, *Facebook and Cambridge Analytica: What You Need To Know as Fallout Widens*, The New York Times, 2018; https://www.nytimes.com/2018/03/19/technology/ facebook-cambridge-analytica-explained.html; accessed 10 February 2022.

G. Greenleaf, *The Influence of European Data Privacy Standards Outside Europe: Implications for Globalization of Convention 108*, International Data Privacy Law, 2 (2), 2012, pp. 68–92. https://doi.org/10.1093/idpl/ips006

J. Halliday, *Twitter's Tony Wang: "We Are the Free Speech Wing of the Free Speech Party,* The Guardian, 2012; https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech; accessed 10 February 2022.

D. Hardegger, *A First Holistic "4th Space" Concept*, Proceedings, 81(1), 2022; https://doi.org/10.3390/proceedings2022081072

J. Heath, *The Uses and Abuses of Agency Theory*, Business Ethics Quarterly, 19 (4), 2009, pp. 497–528; https://doi.org/10.5840/beq200919430

B. E. Hermalin, M.S. Weisbach, *Transparency and Corporate Governance*, National Bureau Of Economic Research, 2007; http://www.nber.org/papers/w12875

N. P. Hoang, D. Pishva, *Anonymous Communication and Its Importance in Social Networking*, 16th International Conference on Advanced Communication Technology, 2014, pp. 34–39; https://doi.org/10.1109/ICACT.2014.6778917

R. Hofstadter, *The Paranoid Style of American Politics and Other Essays*, Knopf, New York 1965.

C. Hood, *What Happens When Transparency Meets Blame-avoidance?,* Public Management Review, 9 (2), 2007, pp. 191–210; https://doi.org/10.1080/14719030701340275

E. Jardine, *Tor, What Is It Good for? Political Repression and the Use of Online Anonymity-Granting Technologies*, New Media and Society, 20 (2), 2018, pp. 435–452; https://doi.org/10.1177/1461444816639976

M. C. Jensen, H. Meckling, *Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure*, Journal of Financial Economics, 3, 1976, pp. 305–360.

C. Kang, S. Frenkel, *'PizzaGate' Conspiracy Theory Thrives Anew in the TikTok Era*, The New York Times, 2020; https://www.nytimes.com/2020/06/27/technology/pizzagate-justin-bieber-qanon-tiktok.html; accessed 07 March 2022.

J. Kay, *Among the Truthers: A Journey through America's Growing Conspiracist Underground*, HarperCollins, New York 2011.

B. L. Keeley, *Of Conspiracy Theories*, Journal of Philosophy, 96 (3), 1999, pp. 109–126.

D. Kirkpatrick, *The Facebook Effect: The Inside Story of the Company that Is Connecting the World*, Simon & Schuster, New York 2010.

J. T. Klein, *Evaluation of Interdisciplinary and Transdisciplinary Research*, American Journal of Preventive Medicine, 35 (2), 2008, pp. 116–123; https://doi.org/10.1016/j.amepre.2008.05.010

L. Knuttila, *User Unknown: 4chan, Anonymity and Contingency*, First Monday, 16 (10), 2011; https://firstmonday.org/ojs/index.php/fm/article/view/3665/3055

S. Koos, *Forschungsbericht: Die "Querdenker". Wer nimmt an Corona-Protesten teil und warum?* [Research report: The Querdenker. Who attends Corona protests and why?], Universität Konstanz 2021; https://kops.uni-konstanz.de/bitstream/handle/123456789/52497/Koos_2-bnrddxo8opad0.pdf?sequence=1

K. Lee, M. C. Ashton, J. Wiltshire, J.S. Bourdage, B.A. Visser, A. Gallucci, *Sex, Power, and Money: Prediction from the Dark Triad and Honesty–Humility*, European Journal of Personality, 27(2), 2013, pp. 169–184; https://doi.org/10.1002/per.1860

L. Lessig, *The Architecture of Privacy: Remaking Privacy in Cyberspace*, Vanderbilt Journal of Entertainment & Technology Law, 1 (1), 1999, pp. 56–65.

S. Malevich, T. Robertson, *Violence Begetting Violence: An Examination of Extremist Content on Deep Web Social Networks*, First Monday, 3, 2020; https://doi.org/10.5210/fm.v25i3.10421

H. Margetts, *The Internet and Transparency*, Political Quarterly, 82 (4), 2011, pp. 518–521. https://doi.org/10.1111/j.1467-923X.2011.02253.x

A. Meijer, *Understanding Modern Transparency*, International Review of Administrative Sciences, 75 (2), 2009, pp. 255–269; https://doi.org/10.1177/0020852309104175

J. M. Miller, K. L. Saunders, C. E. Farhart, *Conspiracy Endorsement as Motivated Reasoning: The Moderating Roles of Political Knowledge and Trust*, American Journal of Political Science, 60 (4), 2016, pp. 824–844; https://doi.org/10.1111/ajps.12234

A. Morisson, *A Typology of Places in the Knowledge Economy: Towards the Fourth Place*, in: New Metropolitan Perspectives. ISHT 2018. Smart Innovation, Systems and Technologies, F. Calabrò, L. Della Spina, C. Bevilacqua (eds.), Springer, Cham, 2019, pp. 444–451; https://doi.org/10.1007/978-3-319-92099-3_50

A. Mungiu-Pippidi, *The Quest for Good Governance*, Cambridge University Press, Cambridge, 2015; https://doi.org/10.1017/CBO9781316286937

A. Mungiu-Pippidi, R. Dadašov, *Measuring Control of Corruption by a New Index of Public Integrity*, European Journal on Criminal Policy and Research, 22(3), 2016, pp. 415–438; https://doi.org/10.1007/s10610-016-9324-z

A. Mungiu, *Corruption: Diagnosis and Treatment*, Journal of Democracy, 17 (3), 2006, pp. 86–99. https://doi.org/10.1353/jod.2006.0050

K. Murphy, *In Online Dating, 'Sextortion' and Scams*, The New York Times, 2016; https://www.nytimes.com/2016/01/17/sunday-review/in-online-dating-sextortion-and-scams.html?searchResultPosition=11; accessed 01 March 2022.

OCCRP, Daraj, Süddeutsche Zeitung, NDR, *Historic Leak of Swiss Banking Records Reveals Unsavory Clients*, Suisse Secrets, 2022; https://www.occrp.org/en/suisse-secrets/ historic-leak-of-swiss-banking-records-reveals-unsavory-clients; accessed 02 March 2022.

A. Ohlheiser, *"We Actually Elected a Meme as President": How 4chan Celebrated Trump's Victory*, The Washington Post, 2016; https://www.washingtonpost.com/news/the-intersect/wp/2016/11/09/we-actually-elected-a-meme-as-president-how-4chan-celebrated-trumps-victory/; accessed 02 March 2022.

R. Oldenburg, *The great Good Place: Café, Coffee Shops, Community Centers, Beauty Parlors, General Stores, Bars, Hangouts, and How They Get You Through the Day*, Paragon House, New York 1989.

Y. J. Park, *Digital Literacy and Privacy Behavior Online*, Communication Research, 40 (2), 2013, pp. 215–236; https://doi.org/10.1177/0093650211418338

R. D. Putnam, *Bowling Alone: America's Declining Social Capital*, Free Press, 2000; https://doi.org/10.4324/9780203805749

A. Rabin-Havt, Media Matters, *Lies, Incorporated: The World of Post-Truth Politics*, Anchor Books, New York 2016.

A. Rege, *What's Love Got to Do with It? Exploring Online Dating Scams and Identity Fraud*, International Journal of Cyber Criminology, 3(2), 2009, pp. 494–512; http://

ra.ocls.ca/ra/login.aspx?url=http://search.ebscohost.com/login.aspx?direct=true&db=i3h&AN=59256420&site=eds-live

T. Rid, *Active Measures: The Secret History of Disinformation & Political Warfare*, Profile Books, London 2020.

P. Robinson, *Flying Blind: The 737 MAX Tragedy and the Fall of Boeing*, Doubleday, New York 2021.

T. Sardá, *The Dark Side of the Internet: A Study about Representations of the Deep Web and the Tor Network in the British Press*, Doctoral thesis, Loughborough University, Loughborough 2020.

K. Starbird, L. Palen, *Pass It On?: Retweeting in Mass Emergency*, Proceedings of the 7th International ISCRAM Conference, 2010; http://idl.iscram.org/files/starbird/2010/970_Starbird+Palen2010.pdf

J.E. Stiglitz, *Transparency and Government*, in: The Right to Tell: The Role of Mass Media in Economic Development, R. Islam, S. Djankov, C. McLiesh (eds.), The World Bank, Washington, DC 2002, pp. 27–44.

O. Storbeck, *BaFin Boss "Believed" Wirecard Was Victim until Near the Ned*, Financial Times, 2021; https://www.ft.com/content/a021012e-bd2e-44d5-a160-96d997c662f1; accessed 03 March 2022.

C. R. Sunstein, A. Vermeule, *Conspiracy Theories*, University of Chicago Public Law & Legal Theory Working Paper, No. 199, 2008; http://ssrn.com/abstract=1084585

C. S. Taber, D. Cann, S. Kucsova, *The Motivated Processing of Political Arguments*, Political Behavior, 31 (2), 2009, pp. 137–155; https://doi.org/10.1007/s11109-008-9075-8

Y. Theocharis, A. Cardenal, S. Jin, T. Aalberg, D.N. Hopmann, J. Strömbäck, L. Castro, F. Esser, P. Van Aelst, C. de Vreese, N. Corbu, K. Koc-Michalska, J. Matthes, C. Schemer, T. Sheafer, S. Splendore, J. Stanyer, A. Stępińska, V. Štětka, *Does the Platform Matter? Social Media and COVID-19 Conspiracy Theory Beliefs in 17 Countries*, New Media & Society, 2021; https://doi.org/10.1177/14614448211045666

A. Thompson, *The Measure of Hate on 4Chan*, Rolling Stone, 2018, https://www.rollingstone.com/politics/politics-news/the-measure-of-hate-on-4chan-627922/; accessed 02 March 2022.

R. Thomson, N. Ito, H. Suda, F. Lin, Y. Liu, R. Hayasaka, R. Isochi, Z. Wang, *Trusting Tweets: The Fukushima Disaster and Information Source Credibility on Twitter*, ISCRAM 2012 Conference Proceedings — 9th International Conference on Information Systems for Crisis Response and Management, 2012; https://www.emknowledge.org.au/ISCRAM2012/ proceedings/112.pdf

Transparency International, 25 Corruption Scandals that shook the world, 2019, https://www.transparency.org/en/news/25-corruption-scandals; accessed 02 March 2019.

Z. Tufekci, *Twitter and Tear Gas: The Power and Fragility of Networked Protest*, Yale University Press, New Haven 2017.

M. Tuters, E. Jokubauskaitė, D. Bach, *Post-Truth Protest: How 4chan Cooked Up the Pizzagate Bullshit*, M/C Journal, 21(3), 2018, pp. 1–18; https://doi.org/10.5204/mcj.1422

Twitter, *About Verified Accounts*, n.d.; https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts; accessed 04 March 2022.

Twitter, *Ban Evasion Policy*, 2020; https://help.twitter.com/en/rules-and-policies/ban-evasion; accessed 04 March 2022.

J. E. Uscinski, A.M. Enders, C. Klofstad, M. Seelig, J. Funchion, C. Everett, S. Wuchty, K. Premaratne, M. Murthi, *Why Do People Believe COVID-19 Conspiracy Theories?*, Harvard Kennedy School Misinformation Review, 1 (April), 2020; https://doi.org/10.37016/mr-2020-015

J. E. Uscinski, J. M. Parent, *American Conspiracy Theories*, Oxford University Press, Oxford 2014.

S. Vaidhyanathan, *Anti-social Media: How Facebook Disconnects Us and Undermines Democracy*, Oxford University Press, Oxford 2018.

P. Van Aelst, J. Strömbäck, T. Aalberg, F. Esser, C. de Vreese, J. Matthes, D. Hopmann, S. Salgado, N. Hubé, A. Stępińska, S. Papathanassopoulos, R. Berganza, G. Legnante, C. Reinemann, T. Sheafer, J. Stanyer, *Political Communication in a High-choice Media Environment: a challenge for democracy*, Annals of the International Communication Association, 41 (1), 2017, 3–27; https://doi.org/10.1080/23808985.2017.1288551

G. J. van Hardeveld, C. Webber, K. O'Hara, *Deviating From the Cybercriminal Script: Exploring Tools of Anonymity (Mis)Used by Carders on Cryptomarkets*, American Behavioral Scientist, 61 (11), 2017, pp. 1244–1266; https://doi.org/10.1177/ 0002764217734271

J. W. van Prooijen, K. M. Douglas, *Conspiracy Theories as Part of History: The Role of Societal Crisis Situations*, Memory Studies, 10 (3), 2017, pp. 323–333. https://doi.org/ 10.1177/1750698017701615

L. Verdoliva, *Media Forensics and DeepFakes: An Overview*, IEEE Journal on Selected Topics in Signal Processing, 14 (5), 2020, pp. 910–932; https://doi.org/10.1109/ JSTSP.2020.3002101

T. Wang, *Talking to Strangers: Chinese Youth and Social Media*, Doctoral Thesis, University of California, San Diego 2013.

M. Wendling, *The Saga of "Pizzagate": The Fake Story that Shows How Conspiracy Theories Spread*, BBC News, 2016; http://www.bbc.com/news/blogs-trending-38156985; accessed 03 March 2022.

R. White, *A Third Place*, New Zealand Geographic, 2018; https://www.nzgeo.com/stories/a-third-place/; accessed 03 March 2022.

M. T. Whitty, T. Buchanan, *The Online Dating Romance Scam: The Psychological Impact on Victims — Both Financial and Non-financial*, Criminology and Criminal Justice, 16 (2), 2016, pp. 176–194; https://doi.org/10.1177/1748895815603773

J. Wideström, A Seeing Place—Connecting Physical and Virtual Spaces, Doctoral Thesis, Chalmers University of Technology, Gothenburg 2020.

O. Winkel, *Electronic cryptography—Chance or threat for modern democracy?*, Bulletin of Science, Technology and Society, 23(3), 2003, 185–191; https://doi.org/10.1177/ 0270467603023003006

S. Zannettou, T. Caulfield, E. De Cristofaro, N. Kourtelris, I. Leontiadis, M. Sirivianos, G. Stringhini, J. Blackburn, *The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources*, Proceedings of the 2017 Internet Measurement Conference, Part F1319, 2017, pp. 405–417; https://doi.org/10.1145/3131365.3131390

A. Zelenkauskaite, P. Toivanen, J. Huhtamäki, K. Valaskivi, *Shades of Hatred Online: 4chan Duplicate Circulation Surge during Hybrid Media Events*, First Monday, 26 (1–4), 2020; https://doi.org/10.5210/fm.v26i1.11075

I. X. Zhang, *Economic Consequences of the Sarbanes-Oxley Act of 2002*, Journal of Accounting and Economics, 44 (1–2), 2007, pp. 74–115; https://doi.org/10.1016/j.jacceco. 2007.02.002

ABOUT THE AUTHORS:

Christoph M. Abels — Doctoral Researcher, Hertie School, Friedrichstraße 180, 10117 Berlin, Germany.

Email: c.abels@phd.hertie-school.org

Daniel Hardegger — PhD, Research Fellow, ZHAW School of Management and Law, Gertrudstrasse 15, 8401 Winterthur, Switzerland.

Email: daniel@hardegger.eu

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Dustin Gray

# MODERN FORMS OF SURVEILLANCE AND CONTROL

doi: 10.37240/FiN.2022.10.zs.9

***ABSTRACT***

In todays advanced society, there is rising concern for data privacy and the diminution thereof on the internet. I argue from the position that for one to enjoy privacy, one must be able to effectively exercise autonomous action. I offer in this paper a survey of the many ways in which persons autonomy is severely limited due to a variety of privacy invasions that come not only through the use of modern technological apparatuses, but as well simply by existing in an advanced technological society. I conclude that regarding the majority of persons whose privacy is violated, such a violations are actually initiated and upheld by the users of modern technology themselves, and that ultimately, most disruptions of privacy that occur are self-levied.

**Keywords**: philosophy of technology, data privacy, surveillance, autonomy.

## 1. INTRODUCTION

How much should we care about our right to privacy, and how much of a role does it play in the total amount of autonomy we experience? Does it make sense to believe that "privacy is a function of liberty" as some do (Rusbridger, MacAskill, 2014)? If we are to follow this line of reasoning, then we are bound to the presupposition that to experience liberty, we must also *have the option* to keep as much of our lives private as we deem appropriate. In doing so, we would be living by a specific self-determined rule and to that extent, have autonomy.

However, an important consideration to make regarding autonomy and privacy is that in virtue of having the former, the rational agent has the final say on how highly she values the latter. The mere exercise of choice as to whether one's privacy is important or not is in itself emblematic of autonomous action. The argument I want to make, however, is that the invasion of privacy that occurs by means of what I call *technological surveillance*—as administered to everyone who exists in today's advanced society—can be

regarded as impermissible. "Technological surveillance" should be recognized in its use throughout the paper as the unwarranted audio, visual, and or digital monitoring of a rational agent's affairs by another.

In this paper, I will aim to provide a greater understanding of what it means to have our privacy pilfered by means of surveillance in a variety of capacities. There will be a discussion on how surveillance is used to control persons within a given society and how this can be seen as a form of oppression that is—in many ways—*self-instituted*. I will argue that many seem to, without concern, place themselves in a position to be regulated in this manner. While some may be oblivious, others simply remain indifferent in regard to the numerous structures put in place to ensure that residents of this and many other countries are being watched, listened to, and otherwise monitored every day (Schwartz, 2017).

To be sure, many methods of surveillance are unavoidable, such as automated license plate readers, public space cameras, and audiovisual surveillance employed on public transportation. I argue, however, that all who use information devices such as mobile phones, computers, and even credit cards *place themselves* in a position to be monitored. Each time these devices are used to make calls, send texts, watch funny cat videos, interact on social media, purchase goods and services, send and receive emails, or conduct internet searches, what is said and heard, sent and received, viewed and posted, bought and sold, and taken interest in is monitored and scrutinized. The use of these devices inherently implies a self-imposed forgoing of one's autonomy. One who places even a shred of value on the retention of her privacy who, in turn, voluntarily discloses her personal information via modern technology could hardly be seen as living by rules set for herself. Not only are these data monitored, but they are stored as well. This retention of another's personal information without permission further demonstrates a loss of autonomy and I argue is deserving of just as much attention as might be given to the manner in which the data is collected. The collection and storage of one's data in this sense does imply a *taking*, but we must not be tempted to think that in collecting and storing our intellectual property it is modern technology that operates as the *taker*. No doubt, we are stripped of our autonomy by technological means, but the identity of the thief lies not in anything technological.

I want to suggest that, ironically, those who most enthusiastically adopt and integrate the modern technological advances that ultimately control them, tend to believe they experience the highest degree of freedom. Furthermore, I argue that the widespread adoption and use of modern technologies is precisely what facilitates the forms of surveillance I am critical of. I will consider those persons who use modern technological devices such as telephones (both standard landline and mobile), computers, "smart" home security systems to be what I call *users*. By integrating the regular use of

these contrivances, persons put themselves in a position to be surveilled by those who I will refer to as *sentinels*. The primary responsibility of the sentinel is to record as much information about the user as possible by means of surveilling her conduct and behavior. But simply monitoring the day-to-day activities of the user will not be enough. Also crucial to the mission of the sentinel is the *storage* of this data for later use, to have a continually growing surplus of information that can be referred back to at any time.

Generally, there are two ways in which the sentinel administers control via surveillance. The first is by way of *corporate* surveillance. The sentinels in this category are technicians and engineers at large and powerful tech companies such as Google, Amazon, and Facebook. The sentinels behind the veil of these entities—as motivated by an all-out perversion of the capitalist venture—have developed an ingenious method to influence the decision-making processes of the consumer. This is done in many ways, but among the most prevalent are the digital monitoring of users' internet searches and the audio surveillance via information devices of what is said by users in their day-to-day lives.

The second means by which users are controlled is what I refer to as *governmental* surveillance. Though specific processes vary, there are three primary methods. The first is simply the audio and visual recording of conduct by means of publicly installed video cameras and microphones. The second is done by the monitoring, recording, and storage of persons' telephone conversations. The third, and possibly most invasive method, is the continuous monitoring and storage of the user's internet activity. In these instances, it turns out that the sentinel is part of the very structure that was originally implemented to protect the rights of its people, yet instead now operates as a system designed to deny that which it promised to protect and uphold.

Notice here that one does not necessarily need to be a user of modern technology to be surveilled. In regard to the first method of governmental surveillance, one only need walk about and congregate in the public arena to become subject to monitoring of this type. This non-user I will refer to more generally as the *citizen*. Being perhaps the greatest minority in existence today, she is still not free from surveillance outside her own home. We might say that all users, too, fall into the category of citizen by existing in an advanced technological society and that one can easily go from user to citizen by way of the use or non-use of modern technology. It is this possibility of transition from user to citizen that implies a choice of degree to which one is controlled. There will be more to come on this toward the end of the paper.

## 2. CORPORATE SURVEILLANCE

So with a general understanding of the ways in which surveillance takes place, I will now move into the specifics of its operation. Let us begin with the corporate method. In her seminal book, *The Age of Surveillance Capitalism*, Shoshana Zuboff gives an extraordinarily detailed account of how corporate surveillance originated and is practiced today. As the title suggests, she argues that surveillance capitalism is the current standard for technological control over the purchasing practices of today's consumers.

> "Surveillance capitalism unilaterally claims human experience as free raw material for translation into behavioral data. Although some of these data are applied to product or service improvement, the rest are declared as proprietary *behavioral surplus*, fed into advanced manufacturing processes known as 'machine intelligence,' and fabricated into *prediction products* that anticipate what you will do now, soon, and later. Finally, these prediction products are traded in a new kind of marketplace for behavioral predictions that I call *behavioral futures markets*. Surveillance capitalists have grown immensely wealthy from these trading operations, for many companies are eager to lay bets on our future behavior." (Zuboff, 2020, p. 8)

From her definition of the term, we find that surveillance capitalism sees the experience of the consumer not as a *subject* to be studied for market research but rather as an *object*. The consumer's experience is considered as data to be compiled as a method by which the corporate sentinel can predict what the user will do next.

Though much could be said about Zuboff's overall analysis, for the purposes of this paper, I will keep a narrow focus on what she discusses concerning the two methods of corporate surveillance listed above: the monitoring of consumers' internet searches and the audio surveillance of consumers' speech. Maintaining that order, let us first explore the ways in which this particular sentinel derives information and makes suggestions based on our internet searches.

Each time you type something into a search engine and press enter, that which you query is captured and stored by, for instance, Google. Zuboff informs us that not only is the keyword itself noted but additionally "each Google search query produces a wake of collateral data such as the number and pattern of search terms, how a query is phrased, spelling, punctuation, dwell times, click patterns, and location" (Zuboff, 2020, p. 67). This collection of information is what Zuboff terms "behavioral data," those data that the user freely provides to Google—or any given search engine—which the sentinel then uses to predict future patterns. Behavioral data alone, though, are of little use to the search provider unless they are *stored*.

During Google's early stages of implementation in the late 90s, "these behavioral by-products were haphazardly stored and operationally ignored" (Zuboff, 2020, p. 67). In the beginning, Google itself did not see the immense potential value of these data; they were merely supplementary bits of information retained within the servers as a result of the users' searches. The original purpose of data collection was, as the company claimed then and still does today, to improve the user's experience by catering search results to the individual based on her search patterns. "Google's engineers soon grasped that the continuous flow of collateral behavioral data could turn the search engine into a recursive learning system that constantly improved search results and spurred product innovations such as spell check, translation, and voice recognition" (Zuboff, 2020, p. 68).

It was not until Google found itself in need of additional revenue streams that behavioral data emerged as a vast untapped mine of profitability. During the first two years of its establishment in 1998, the founders of Google, Larry Page and Sergey Brin, maintained a "passionate and public opposition to advertising" (Zuboff, 2020, p. 74). But in December of 2000, a damning *Wall Street Journal* article incited concerns of future profitability in the company's investors. The article generally targeted many Silicon Valley startups by saying, "Simply displaying the ability to make money will not be enough to remain a major player in the years ahead" (Zuboff, 2020, p. 74). The article maintained that what would be required would be "an ability to show sustained and exponential profits" (Zuboff, 2020, 74). In response to investor anxiety, Page and Brin departed from their earlier convictions on advertising and set the then seven-person internal department, AdWords, on a project to find new streams of revenue. "Operationally, this meant that Google would turn its own growing cache of behavioral data and its computational power and expertise toward the single task of matching ads with queries" (Zuboff, 2020, p. 74). Put simply, the advertising would have to become "relevant" to users. More appropriately, as Zuboff remarks, "a particular ad would be 'targeted' to a particular individual" (Zuboff, 2020, p. 74). She terms this immense reserve of user information as "behavioral surplus." Not only is this what ultimately led to the "sustained and exponential" profits Google was after, it also served as the origin of the epoch of corporate surveillance or what Zuboff would call surveillance capitalism (Zuboff, 2020, p. 99).

Worth noting at this point is an argument made nearly 70 years prior to that of Zuboff's. Martin Heidegger maintained in *The Question Concerning Technology* that the goal of technology is to place that which is derived for modern technological purposing into "standing reserve." "Everywhere everything is ordered to stand by, to be immediately at hand, indeed to stand there just so that it may be on call for a future ordering" (Heidegger, 2013, p. 17). I argue that the collection and storage of user's search patterns on the internet by any means is fundamentally related to this claim.

This brings us sharply to Zuboff's claim that our conduct on the internet is *commodified*. This modern instantiation of human behavior is monitored, commandeered, and stored for the purpose of predicting future instantiations thereof by companies like Google so that they might turn a profit. She claims that what we do online is digitally *dispossessed*.

> "Today's owners of surveillance capital have declared a fourth fictional commodity expropriated from the experiential realities of human beings whose bodies, thoughts, and feelings are as virgin and blameless as nature's once-plentiful meadows and forests before they fell to the market dynamic. In this new logic, *human experience is subjugated to surveillance capitalism's market mechanisms and reborn as 'behavior*.' These behaviors are rendered into data, ready to take their place in a numberless cue that feeds the machines for fabrication into predictions and eventual exchange in the new behavioral futures markets." (Zuboff, 2020, p. 100)

In other words, we ourselves have become the resources mined for standing reserve. "Knowledge, authority, and power rest with surveillance capital, for which we are merely 'human natural resources' " (Zuboff, 2020, p. 100). Those who control the technological powers that be may claim to be monitoring our conduct online to cater their services to our individual wants and needs, but the true motivation has become profitability via appropriation of users' behavioral data (Viadhyanathan, 2011, pp. 21–23).

Another sentinel that has become a leading frontrunner in the use of corporate surveillance is Facebook. Nearly everyone today is aware of the "Like" button. This seemingly harmless digital apparatus is clicked on by Facebook users to express interest in or approval of other users' posts on the social media platform. However, there is a much deeper functionality behind the veil of congeniality proposed by the "Like" button. Each time you "like" a post, something called a "cookie" is installed into your computer, tablet, or smartphone. Not unlike a burrowing parasite, these tiny bits of code embed themselves into your device to establish and allow intersystem communication between Facebook and the end user. The information gained through this exchange is used by Facebook analysts to determine which ads will display based on your interests. Again, the user's behavior online has become a human resource to be exploited for the purpose of targeted advertising that will lead to profitability for the sentinel.

Some might say, however, that this degree of privacy invasion is to be expected. When one signs up for a Facebook account, she is required to read and agree to a lengthy terms and conditions document, which outlines all of this in the privacy section. All Facebook users are informed of the risk they are taking by clicking the "Agree" box. However, in an article published by privacy researcher Arnold Roosendaal, it was found that even non-users of Facebook's services were being monitored as well simply by viewing

webpages associated with Facebook data (Roosendaal, 2010). So as it turns out, even those who do not agree to Facebook's terms are possible targets of corporate surveillance.[1]

Perhaps this, and what was expressed in regard to the data mining tactics employed by Google could be seen as harmless. In fact, there are some who might say they enjoy these predictive features in that they are presented with ads for products they actually are interested in. With these persons, I cannot and will not argue. But I will present one more example that might change the mind of even the most tolerant user.

Zuboff tells of a particularly disturbing service offered by various companies referred to as "service-as-software" (SaaS). She more appropriately deems it as "surveillance as a service" (SVaaS). For example, app-based technologies are being used by financial lenders to monitor the digital and physical behavior of potential borrowers before deciding whether they will provide a loan. One particular app "instantly establishes creditworthiness based on detailed mining of an individual's smartphone and other online behaviors, including texts, emails, GPS coordinates, social media posts, Facebook profiles, retail transactions, and communication patterns" (Zuboff, 2020, p. 172). Not only are these digital data collected, but physical patterns of behavior such as phone charging frequency, whether a user returns calls and how long it takes her to do so, or the distance a user travels each day are also taken into account (Zuboff, 2020, p. 172). Though the common user of information devices might think that data mining for the purpose of targeted advertisement is permissible, this degree of privacy invasion can and will stand directly in the path between a user and her potential to achieve financial security. This instantiation of corporate surveillance entails not the common, "that's just the way it is" mentality. It brings to the forefront a much deeper element of control involved with the surveillance perpetrated by corporate sentinels on users requiring their services.

Thus far, we have explored the actualities of corporate surveillance relating only to the user's conduct online. There is, however, another important feature of this invasive oppressive force that I would like to explore. Much of modern technology today exists in the home, and this is where we its most intimate forms of use occur. Digital assistants such as Alexa and Nest are among the most popular. With these devices, a user can simply verbalize the desire to listen to a particular song or artist, change the temperature on her thermostat, turn lights on and off, lock and unlock doors, etc. These capabilities might seem to provide freedom within one's home but consider also that having these devices installed presupposes the remittance of one's con-

[1] Since Roosendaal's findings, much has transpired. See pages 158–161 of Zuboff's *The Age of Surveillance Capitalism* to learn more about the many allegations made against Facebook regarding its surveillance methods and the ways in which the company defended itself by claiming that these practices were merely a "glitch" or "bug" in the system.

trol to these functionalities. Also worth noting is that many of these devices are actively listening to your speech patterns in search of specific indicators of what you may desire as a consumer. "Pieces of your talk are regularly farmed out in bulk to third-party firms that conduct 'audio review processes' in which virtual scorers, tasked to evaluate the degree of match between the machines text and the original chunk of human speech, review audio recordings retained from smartphones, messaging apps, and digital assistants" (Zuboff, 2020, p. 262). So not only is this data used to provide targeted advertising of goods and services on any device connected to the home system, it is also collected by third party firms to perfect the devices' ability to match what is recorded to the individual user.

It is insisted upon by companies such as Amazon, Google, and Microsoft that these data are anonymous and cannot be linked to individual users, but Zuboff cites the findings of a freelance journalist, A. J. Dellinger, who discovered loopholes in these claims of anonymity.

> "Within the recordings themselves, users willingly surrender personal information—information that is especially valuable in these review processes because they are so specific. Uncommon names, difficult-to-pronounce cities and towns, hyperlocal oddities [...]. I heard people share their full names to indicate a call or offer up location-sensitive information while scheduling a doctor's appointment [...] the recordings capture people saying things they'd never want heard, regardless of anonymity [...]. There isn't much to keep people who are listening to these recordings from sharing them." (Dellinger, 2015)

Zuboff tells of one device in particular that arguably took these capabilities too far. Besides smartphones and digital assistants, Smart TVs are highly sought after by consumers of modern technology. But in 2015, it was found by privacy advocates that Samsung's line of these devices may have been too smart. Not only when instructed to do so, these particular Smart TVs were recording everything said within an earshot of the system. The TVs were capturing phrases such as "*please pass the salt; we're out of laundry detergent; I'm pregnant; let's buy a new car; we're going to the movies now; I have a rare disease; she wants a divorce; he needs a new lunch box; do you love me?*—and sending all that talk to be transcribed by another market leader in voice recognition systems, Nuance Communications" (Zuboff, 2020, p. 263). If we consider the fact that the unique individual fingerprint associated with our voices is something that many firms regard as their sole object of interest as Zuboff has suggested, having the intimate details of our lives recorded in this manner should be alarming at a minimum.

In most cases, I am sensitive to the possible objections that may arise to my claims. But regarding what has been said in this example, I simply will

not concede. Technologies of this nature make possible an inexcusable degree of privacy invasion, and it is my contention that the manner by which these sentinels monitor and store our speech, thoughts, and actions is unquestionably oppressive. We are given no access to check and balance the capabilities of such contrivances, and short of absolute boycott, the oppression will not stop.[2]

As mentioned early on, these instances of corporate surveillance involve the manifestation of an oppressive force that is *self-levied*. We can sit here all day reveling in our accusations that Google, Facebook, and Amazon are wrongfully dispossessing us of our innermost thoughts and feelings, but the truth of the matter is that we are fundamentally the ones to blame, for we, the users, seem unable to live without the various technologies that the sentinels provide. Sure, tech giants such as the ones we have looked at thus far make a convincing case for the necessity to buy what they are selling, and most do. But it must be remembered that in all of this, we do have a choice. And if I am correct, then one will have a difficult time arguing against the oppression imposed by something that one refuses to live without.

## 3. GOVERNMENTAL SURVEILLANCE

It is generally accepted that while in public, our actions and activities are subject to monitoring by both audio and video surveillance equipment. Some of these methods are employed by private companies and some by law enforcement (Gomez, 2019). Some might say that being monitored while in public is just indicative of the world we live in today.[3] It could be argued that the modern advantages associated with existence in a technologically advanced society fundamentally come at the cost of our privacy. But just as we have seen with corporate surveillance, I will show that governmental surveillance is just as—if not more so—oppressive.

Consider the fact that deeply intimate and private aspects of your life are being regularly recorded and stored each time you make a phone call, send an email, or use a search engine. Put simply, when you communicate via telephone or on an internet connected device, *you are being monitored*. But in this case, the deployment of surveillance stems not from capitalist profit motive. In what is to be discussed for the remainder of the paper, I will uncover the aggressive tactics employed by our own government to observe and control its populace.

On October 26, 2011, President George W. Bush signed a piece of legislature known as the USA PATRIOT ACT (Uniting and Strengthening America

---

[2] More will be discussed on this in the conclusion.

[3] Arguably, there is much to be said about audio/visual surveillance of the common citizen, but for the purposes of this project, I adhere mainly to those systems of governmental surveillance involving the monitoring of telephonic and internet communications.

by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism) (USA PATRIOT ACT, 2011). This enabled the National Security Agency (NSA) to monitor and record the phone calls and digital communications of every U.S. citizen.

In June of 2013, a former NSA contractor, Edward Snowden leaked thousands of classified documents to the press revealing the NSA's methods and abilities to intercept all Americans' phone calls and internet traffic (Gellman, Blake, Miller, 2014). Subsequently, President Barack Obama addressed public concerns by describing plans to reform NSA spying. He stated, "They're not abusing authorities in order to listen to your private phone calls, or read your emails" (Ackerman, Roberts, 2014). The original phrasing of the Patriot Act was drafted explicitly in its primary intention to seek out and stop the spread of terrorism. In his speech, President Obama was intending to make the case that the common, law-abiding American need not be concerned and would not be directly affected by the conduct of the NSA.

Upon hearing this speech, one might assume that effective measures would be enacted to protect the privacy of Americans' tele/data communications. However, more recently in 2018, *The New York Times* reported that the NSA had tripled its data collection from U.S. phone companies (Savage, 2018). So though there was a changing of the guard in terms of presidential leadership, the NSA not only continued to monitor residents of the United States but actually increased its efforts in doing so three-fold.

Within the philosophy of technology, there is rising concern for digital privacy and the ethics of data collection. As the emphasis of this paper is on the ethical implications of governmental surveillance and data collection, I call upon our old friend, *utilitarianism*, to better understand the consequences of governmental surveillance and decide whether it can be justified.

Typically, Jeremy Bentham is associated with "act" utilitarianism. An example of such would be a Marine jumping on a hand grenade and thus taking the brunt of its force to ensure the safety of his squad.

> "An action then may be said to be conformable to [the] principle of utility, or, for shortness sake, to utility, (meaning with respect to the community at large) when the tendency it has to augment the happiness of the community is greater than any it has to diminish it." (Bentham, 2000, p. 15)

For Bentham, an act is good when its consequences increase the happiness of the community at large. In following the language used by Bentham and the broader logic of language, we could—at the very least, generally—call the American public a community.

In slight variation, John Stuart Mill brought about what is commonly known as "rule" utilitarianism. An example of this would be a given company's policy that if an employee is feeling ill that she not come into the office,

for to do so would create the possibility of getting others sick. "All action is for the sake of some end, and rules of action, it seems natural to suppose, must take their whole character and colour from the end to which they are subservient" (Mill, 2001, 6). Mill suggests that rather than actions, we should focus on which rules will promote the highest degree of happiness for those who fall subject to them.

How might we apply these variations of utilitarianism to the Patriot Act considering that though it was ostensibly put in place to protect all Americans from the threat of terrorist infiltration and attack, it also necessitates the unwarranted audio and digital surveillance of all American citizens? The Patriot Act operates as a piece of legislation that involves specific circumstances and persons. By its own language, we are led to believe that the intended targets of surveillance are those suspected to be involved with terrorist organizations and capable of committing acts of terrorism upon innocent civilians. However, as has been shown, the focus is not centralized in this manner. *All* Americans must be monitored in order to weed out those that might pose a threat. As a matter of policy, it is a matter of rule. The NSA has made the implicit claim that *as a rule*, it should retain the ability to monitor everyone in search of radical terrorists. Framed this way, I am inclined to think that what we are dealing with is rule utilitarianism, at least *prima facie*. The aim of the Patriot Act may very well be to protect the lives of the American people, but I argue that it carries with it the consequence of innocent Americans being monitored in a way that limits their autonomy. It denies the right to privacy of those it is supposed to protect.

Whether viewed as action or rule, one could argue that the consequences of the Patriot Act do promote the greatest degree of happiness or pleasure—or in this case, security—for the majority of those impacted. An advocate of this variety could take the stance that if her autonomy must be limited by monitoring her phone calls and internet traffic in order to gain protection from terrorist threat, so be it. Besides, she has nothing to hide, right? For this particular user, the ends justify the means.

In support of utilitarianism, Peter Singer offers a formulation that attempts to ameliorate both of the accounts previously mentioned. He suggests that when making any ethical decision, we must take ourselves out of the picture. We must consider it as applying to everyone collectively and, in so doing, we must never allow our specific individual desires to influence or intrude upon this process. "In accepting that ethical judgments must be made from a universal point of view, I am accepting that my own interests cannot, simply because they are my interests, count more than the interests of anyone else" [Singer 1979, 12]. Singer argues that whether we are looking at acts or rules, we must consider the consequences for those impacted *above and beyond* our motivation for their creation. Let us look at the issue from this perspective and see what comes about.

One could clearly speculate ulterior motives, but for the moment, I will grant that the singular motive behind the creation and implementation of the Patriot Act was to identify terrorist threats via telephonic and internet surveillance. Those involved in the creation and execution of the Patriot Act—the NSA and the U.S. federal government—enjoy the benefit not only of having unfettered access to all Americans' tele/data communications and patterns of online conduct, but they also have the benefit of referring back to any specific data of their choosing as all that is monitored is stored. This is an actual consequence of the actions allowed by the Patriot Act. With this in mind, recall that the aim of the Patriot Act is to identify terrorist threats, and the method is mass surveillance of all persons in this country. The employment of this process certainly makes possible the identification of terrorists, for if you are watching everyone all the time, the chances that you will be able to locate the bad apples are good. Speaking literally, this is how bad apples are found. From this, we can correctly surmise that dragnet governmental surveillance can amount to the possibility of identifying terrorist threats, but what can we say of *actual* discovery?

On June 18, 2013, NSA Director General Keith Alexander testified before the U.S. House Select Intelligence Committee that governmental surveillance programs authorized by the Patriot Act "had helped prevent 'potential terrorist events over 50 times since 9/11'" (Nakashima, 2013). Though by their very description, these events were characterized as being merely potential, their identification did, in fact, seem to be actual. On October 16, 2013, it was reported that Alexander would be stepping down as NSA Director. This likely came in the wake of Snowden's exposing the agency's indiscriminate sweeping surveillance of American's telephone and internet data. It is also likely that Alexander's resignation came as a result of his admission that the actual number of potential terrorist events was an over exaggeration (Live Leak, 2020). Though the number of terrorist threats identified via governmental surveillance programs turned out to be lower than Alexander's original declaration, we could grant that at least some degree of terrorist threat was actually identified. In making an argument for utility, however, we must consider the *entire* scope of consequence.

Besides the consequence of identifying terrorist threats, I have demonstrated another that comes in the form of widespread and indiscriminate surveillance of American's telephone calls and their conduct online. Returning to the question concerning utilitarianism posed earlier, let us not think in terms of pain or pleasure, but rather in those of security and risk. I argue that ubiquitous governmental surveillance authorized by the Patriot Act does not follow an act model of utilitarianism. This is because the act does not promote a higher degree of security than is justified to eliminate risk of terrorist attack. We could imagine such adherence only if it were the case that once identified as a terrorist threat—by having compelling reasons to

believe so—surveillance was then implemented to gain further intelligence. Only surveillance of *known* terrorist threats would meet the necessary conditions of act utilitarianism. The individual act of surveillance would be permissible because the ends would justify the means.

Can we then say that governmental surveillance meets the conditions necessary to conform to the precedent of rule utilitarianism? Well, considering that the overarching and indiscriminate surveillance taking place as I type these very words does operate as a rule, we might be inclined to think so. But when we consider that all Americans—innocent or otherwise—as well as possible terrorist organizations are targeted, the methodology attracts more intuitive scrutiny. Surveillance on a scale this massive creates a situation in which the entire civilian population enjoys a disproportionately lower level of benefit than is promised by the means. Therefore, it is not clear that governmental surveillance can be justified under a rule model of utilitarianism. It is not clear that the level of security promised justifies the degree of privacy relinquishment required to fulfill it.

Finally, consider that the monitoring of private affairs and especially the retention of collected data involves the unabashed denial of Americans' 4th Amendment right to be secure in their persons, houses, papers, and effects. Governmental monitoring, collection, and storage of telephone call transcriptions and internet traffic equates simply to illegal search and seizure of one's intellectual property. Considering this, it seems that even outside the scope of utilitarianism governmental surveillance entails a legitimate violation of rights that are supposed to be guaranteed by those laid out in the U.S. constitution. Whether it is viewed under a consequentialist lens or simply considered using general ethical reasoning, I argue that surveillance of this nature is both unwarranted and unjustified.

I have also suggested that surveillance of this nature involves a loss of autonomy suffered by anyone who uses a telephone or computer, which turns out to be a vast majority of persons in this country. Again, we can presume the objection will be made that if one has nothing to hide, then surveillance of this kind is of no consequence and, therefore, poses no threat to one's autonomy. I will, however, ask this brand of objector to consider the way she conducts herself in private as opposed to in public. Before a date, many try on a number of outfits in private for the sole purpose of selecting the only one they want to be seen in by their partner in public. Those who tremble in fear at the mere idea of singing a song in front of an audience might do so emphatically in the shower alone. It is no secret that many people "pleasure themselves" sexually on a regular basis and feel there should be no stigma attached to such a practice as it serves as a healthy method of satisfying one's urges and relieves stress. Would such a person feel comfortable doing this in front of a group of NSA agents? I wager not.

The point here is that there are any number of strange and normal things we do in private *because* we are in private. An actual consequence of the Patriot Act is that one has to consider that she is being monitored as she researches birth control methods, seeks out divorce lawyers, and diagnoses strange rashes online. These intimate affairs are ones I am inclined to think that most would wish to remain private, but the Patriot Act removes the possibility for privacy in such conduct and in so doing disallows the possibility of one's retention of autonomy. In considering these autonomy limiting factors in conjunction with the utilitarian analysis provided above and the fact that this policy effectively authorizes unlawful search and seizure on a blindingly massive scale, I argue that the Patriot Act and its subsequent authorization of NSA spying on innocent civilians follows no principle of utility or morality whatsoever.

## 4. SURVEILLANCE AS A FORM OF CONTROL

For those who cherish our constitutionally guaranteed right to privacy, much of what I have said here is troubling. Of those who contend that NSA surveillance is unproblematic in that they "have nothing to hide," we might ask why they have blinds in their windows or doors on their bathrooms. We might ask if they are aware of the NSA's surveillance of pornography viewing habits, would they draw the same conclusion (Greenwald, Grim, Gallagher, 2017)? In deciding how to respond to the implications of NSA surveillance, I offer the words of philosopher Robert Paul Wolff as cited by Singer:

> "The defining mark of the state is authority, the right to rule. The primary obligation of man is autonomy, the refusal to be ruled. It would seem, then, that there can be no resolution of the conflict between the autonomy of the individual and the putative authority of the state. Insofar as a man fulfills his obligation to make himself the author of his decisions, he will resist the state's claim to have authority over him." (Singer, 1979, p. 293)

The point Wolff is making here is that inherently, the state and its people will always be at an impasse due simply to his declaration that the state demands authority and its citizens demand autonomy. What all of this really amounts to is *control*. Governmental surveillance is nothing more than the latest technological method to ensure that control of its citizens remain in the hand of the state. It is no secret that we civilians vastly outnumber the total amount of both police officers and military, yet government officials fear not any uprising or power shift of any kind. This is because shrewdly they have taken control by technological means to ensure that the teenagers will never throw a party because the parents will never leave town.

As far as the use of modern technology, however, I fear that the convictions expressed by Wolff have gone the way of the buffalo. In a society so infatuated with modern technology, its residents have become convinced—whether they know it or not—that unwavering adherence to the rules decreed by another are acceptable under any conditions, even when they remove the ability to live by those we might give ourselves.

As users of modern technology, we have voluntarily succumbed to the allure of modern digital existence. It is unlikely that many users would even consider the possibility of being what I referred to in the beginning of the paper as a mere citizen. There may be those rare few who refuse to participate, and to them I am more or less in accord. But for the masses—for that overwhelmingly disproportionate majority of persons who make the ritualistic use of modern technology requisite for their daily patterns of existence—there is no freedom from the bondage of corporate nor governmental surveillance.

**REFERENCES**


J. Bentham, Laurence J. Lafleur, *An Introduction to the Principles of Morals and Legislation*, Batoche Books, Kitchener, Ontario 2000.

A. J. Dellinger, *I Took a Job Listening to Your Siri Conversations*, *Daily Dot*, March 25, 2015; accessed May 28, 2020; https://www.dailydot.com/debug/siri-google-now-cortana-conversations/

B. Gellman, Blake A., Miller G., *Edward Snowden Comes Forward as Source of NSA Leaks*, The Washington Post, June 9, 2013; https://www.washingtonpost.com/politicsintelligence-leaders-push-back-on-leakers-media/2013/06/09/fff80160-d122-11e2-a73e-826d299ff459_story.html

L. Gomez, *Cameras on Nearly 3,000 Street Lights all over San Diego, Police Take Interest in Video*, The San Diego Union-Tribune, March 19, 2019; https://www.sandiegouniont ribune.com/opinion/the-conversation/sd-san-diego-street-light-sensors-camera-for-law-enforcement-use-20190319-htmlstory.html

G. Greenwald, Grim R., Gallagher R., *Top-Secret Document Reveals NSA Spied on Porn Habits as Part of Plan to Discredit 'Radicalizers'*, Huffington Post, Updated December 6, 2017; https://www.huffpost.com/entry/nsa-porn-muslims_n_4346128?13855 26024=

M. Heidegger, *The Question Concerning Technology and Other Essays*, William Lovitt (trans.), Harper Perennial, New York 2013, 17.

L. Leak, *Snoop Flies Coop: NSA Head to Quit After Lying, Failing to Explain Spy Overreach*, LiveLeak.com; accessed June 23, 2020; https://www.liveleak.com/view?i=2b8_1382134733

Mill, John Stewart. *Utilitarianism,* Batoche Books, Kitchener, Ontario 2001.

A. Roosendaal, *Facebook Tracks Everyone: Like This!* Tilburg Law School Legal Studies Research Paper Series, 3 (November), 2010, 1–10; https://papers.ssrn.com/sol3/papers. cfm?abstract_id=1717563

A. Rusbridger, Ewen MacAskill, *Edward Snowden Interview—The Edited Transcript*, *The Guardian*, July 18, 2014; https://www.theguardian.com/world/2014/jul/18/-sp-edward-snowden-nsa-whistleblower-interview-transcript

Ch. Savage, "N.S.A., *Triples Collection of Data from U.S. Phone Companies*, The New York Times, May 4, 2018; https://www.nytimes.com/2018/05/04/us/politics/nsa-surveillance-2017-annual-report.html

S. Schwartz, *9 Ways You're Being Spied on Every Day*, Huffington Post, updated December 6, 2017. https://www.huffpost.com/entry/government-surveillance_n_5084623? utm_campaign=share_email&ncid=other_email_063gt2jcad4

P. Singer, *Practical Ethics*, Cambridge University Press, Cambridge, England 1979.
A. Spencer, D. Roberts, *Obama Presents NSA Reforms with Plan to End Government Storage of Call Data*, The Guardian, January 17, 2014; https://www.theguardian.com/world/2014/jan/17/obama-nsa-reforms-end-storage-americans-call-data
U.S. Government Printing Office, *USA PATRIOT ACT*, accessed March 27, 2022; https://www.govinfo.gov/content/pkg/PLAW-107publ56/html/PLAW-107publ56.htm
S. Vaidhyanathan, *The Googlization of Everything: (and Why We Should Worry)*, University of California Press, Berkeley, 2011; http://ebookcentral.proquest.com/lib/ucsc/detail.action?docID=656365.
S. Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Public Affairs, New York 2020.

ABOUT THE AUTHOR — Master of Arts, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, U.S.
Email: dugrayucsc@protonmail.com

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Magnus Johnsson

# PERCEPTION, IMAGERY, MEMORY AND CONSCIOUSNESS

doi: 10.37240/FiN.2022.10.zs.10

***ABSTRACT***

I propose and discuss some principles that I believe are substantial for perception, various kinds of memory, expectations and the capacity for imagination in the mammal brain, as well as for the design of a biologically inspired artificial cognitive architecture. I also suggest why these same principles could explain our ability to represent novel concepts and imagine non-existing and perhaps impossible objects, while there are still limits to what we can imagine and think about. Some ideas regarding how these principles could be relevant for an autonomous agent to become functionally conscious are discussed as well.

**Keywords**: perception, memory, expectations, imagination, consciousness, self-organization, feature maps, associative learning, multimodal integration, cognitive architecture.

## 1. INTRODUCTION

The astonishing abilities of the mammalian brain raise the question of the principles by which it is organized. Since it is evolved rather than designed, such principles should be simple rather than complicated. This seems to be contradicted by the brain's remarkably advanced abilities. I believe that this contradiction is false and that the advanced capabilities of the brain are indeed based on fairly simple principles, but which are reused over and over again at different levels of complexity.

Below I explain and motivate why I think rather simple principles regarding self-organization; internal self-supervision; and associatively, hierarchically, recurrently connected topology preserving feature representations that reflect the probability distribution of their input are important in the mammal brain as well as how that insight can be used artificially.

Simple principles, I believe, employed over and over again by nature at various levels of complexity, are behind astonishingly complex abilities, such as perception, imagery, and functional consciousness in the mammal

brain. The same principles can explain why we sometimes tend to perceive our expectations rather than what is really out there; how we construct and fill in the gaps in our perceptions within and between various sensory modalities when the sensory input is limited; multimodal integration. How various memory systems, imagery and perception fit together and can be explained by the same principles.

I will also discuss how the corresponding faculties could be implemented in an artificial biologically inspired cognitive architecture by employing the same principles in the same way as has presumably been done through evolution by nature. By looking at how the mammal brain is structured and by identifying its crucial components and how these are interconnected, knowledge can be obtained that together with the identified principles enables a systems level approach to modeling perception as well as the integration of various sensory modalities, memory, imagery, the generation of an inner world and functional consciousness in a biologically inspired cognitive architecture modeled on the mammal brain.

This systems level modeling approach means that, though not modeling crucial components and their interconnections in detail, general principles also adhered to by their biological counterparts should be identified and followed in the design of the cognitive architecture.

The components' functionality can be implemented with mechanisms that model the systems at a suitable level of accuracy. Later they can be re-implemented by other mechanisms for accuracy and performance reasons, or if more efficient implementations are found.

We could go about working on a bio-inspired systems-level cognitive architecture in various ways. At one extreme, we could work from a more holistic starting point by identifying crucial components and interactions found in the neural systems of biological organisms. Then we could implement maximally simplified versions of these and try to make them work together as well as possible. Examples of components in such an architecture inspired by a mammal brain could be a maximally simplified visual system and a maximally simplified mechanism, or set of mechanisms, corresponding to the Basal ganglia etc. Inspired by the work of Valentino Braitenberg (Braitenberg, 1984) and the robotics physicist Mark W. Tilden, I believe such simplified but complete cognitive architectures would still enact interesting behaviors.

At the other extreme, we could work on individual components while trying to optimize these to perform at a human level or beyond. Many artificial perception researchers work at this extreme, e.g. by creating computer vision systems that in some respects even exceed the abilities of humans.

My approach is somewhere in the middle. I try to figure out general principles for not necessarily complete, but more composed architectures at an intermediary level. Hence my focus is not whether component imple-

mentations are optimized for performance. Following a systems level approach, individual components can be reimplemented iteratively at later stages for performance, accuracy or for other reasons, but this is not the focus here. Thus, the work on general principles can be isolated from the engineering questions of performance.

The perceptual parts of a cognitive architecture built according to these ideas employ to a large extent self-organizing topographical feature representations. Such feature representations are somewhat reminiscent of what has been found in mammal brains. These topographical representations are connected hierarchically, associatively and recurrently. Hierarchies of increasingly complex feature representations—and in the extension different architectural components, possibly distributed—self-organize while supervising each other's associative learning/adaptation over space (by associative connections) and over time (by recurrent associative connections). For example, such an architecture contains hierarchies of topographically ordered feature representations within sensory submodalities. To an extent, these hierarchies also cross the borders of different sensory submodalities, and even the borders of different sensory modalities. The topographically ordered feature representations connect associatively at various hierarchical levels within, but also across, sensory submodalities, modalities and to systems outside the perceptual parts, e.g. motor representations.

Below I discuss some principles that I believe are substantial to perception, various kinds of memory, expectations and imagery in the mammal brain and for the design of a bio-inspired artificial cognitive architecture. I also suggest why these principles could explain our ability to represent novel concepts and imagine non-existing and perhaps impossible objects, while there are still limits to what we can imagine and think about.

I will also present some ideas regarding how these principles could be relevant for an autonomous agent to become p-conscious (Block, 1995) in the sense defined by Bołtuć (2009), i.e. as referring to first-person functional awareness of phenomenal information. Whether such an autonomous agent would also be conscious in a non-functional first-person phenomenological sense, i.e. h-conscious, adopting again the terminology of Bołtuć (2009), and thus experience qualia of its own subjective first-person experiences of external objects and inner states, is another matter. The latter question belongs to the hard problem of consciousness (Chalmers, 2003). The difficulty with that problem is that a physical explanation in terms of brain processes is an explanation in terms of structure and function, which can explain how a system's behavior is produced, but it is harder to see why the brain processes are accompanied by subjective awareness of qualia. According to Chalmers (2003) all metaphysical views on phenomenal consciousness are either reductive or nonreductive, and he considers the latter to be more promising. Nonreductive views require a re-conception

of physical ontology. I suggest that the bio-inspired principles proposed in this paper have relevance for p-consciousness. Hence a cognitive architecture employing these ideas would probably become at least p-conscious. However, it is possible that h-consciousness is not a computational process, and I will not take a final position on the issue of phenomenal h-consciousness in this paper.

## 2. FEATURE REPRESENTATIONS

Topographically ordered maps are inherent parts of the human brain. There are continuously ordered representations of receptive surfaces across various sensory modalities, e.g. in the somatosensory and visual (van Essen, 1985) areas, in neuron nuclei and in the cerebellum.

The size of the representational area in such ordered representations depends on the behavioral importance and frequency of the represented input. For example, the representation of the fovea is much larger than the rest of the retina, and the representation of the fingertip is proportionally larger than the rest of the finger.

There are also more abstract topographically ordered representations in the brain, e.g. frequency preserving tonotopic maps (Tunturi, 1950; 1952; Reale, Imig, 1980) in primary auditory areas, and color maps in V4 (Zeki, 1980) in the visual areas.

In a model of such self-organized topographically ordered representations, essential relations among data should be made explicit. This could be achieved by forming spatial maps at an appropriate abstraction level depending on the purpose of the model. For a reasonable computational efficiency, the focus should be on main properties without any accurate replication of details. Reasonable candidates for a basic model corresponding to a topographically ordered representation in the brain satisfying these conditions are the Self-Organizing Map, SOM (Kohonen, 1988) and its variants. Such a basic model forms a fundamental building block—not to be confused with the crucial components discussed above—in the perceptual parts of a bio-inspired cognitive architecture.

Examples of suitable candidates, beside the SOM, are the Growing Grid (Fritzke, 1995) and the Growing Cell Structure (Fritzke, 1994). In addition to the adaptation of the neurons, these models also find suitable network structures and topologies through self-organizing processes. Other examples are the Tensor-Multiple Peak SOM, T-MPSOM (Johnsson et al., 2006) or the Associative Self-Organizing Map (Johnsson et al., 2009). The latter, or rather the principles it instantiates, are crucial for the principles of the perceptual parts of a cognitive architecture discussed in this paper and will be elaborated on below.

The SOM develops a representation that reflects the distance relations of the input, which is characteristic of lower levels of perception. If trained with a representative set of input, the SOM self-organizes into a dimensionality reduced and discretized topographically ordered feature representation also mirroring the probability distribution of received input. The latter means that frequent types of input will be represented with better resolution in the SOM. This corresponds to, for example, the development of a larger representational area of the fingertip than the rest of the finger in the brain, which was discussed above. Hence the SOM is reminiscent of the topographically ordered representations found in mammalian brains.

In a sense, the topographically ordered map generated by a SOM—and in the extension an interconnected system of SOMs—is a conceptual space (Gärdenfors, 2000) generated from the training data through a self-organizing process.

Due to the topology-preserving property of the SOM similar input elicit similar activity, which provides systems based on the SOM with an ability to generalize to novel input.

A SOM can be trained to represent various kinds of feature, including phenomenal ones. The latter would turn the SOM into a phenomenal content map (Damasio, 2010). For example, a SOM can be trained to represent directions of lines/contours (as in V1), colors (as in V4), or more complex features such as the postures and gesture movements of an observed agent (Buonamente et al., 2016), or the words of a text corpus ordered in a way that reflects their semantic relations (Ritter, Kohonen, 1989). Employing SOMs or other topographically ordered feature representations to represent phenomenal features together with the general design principles for a bio-inspired cognitive architecture suggested in this paper, would enable strong semantic computing (Bołtuć, 2018).

Other models of a self-organizing topology preserving feature representation are possible and might turn out to be more suitable for various reasons such as performance and accuracy. However, as also mentioned above, that is beyond the point of this paper, which aims at presenting higher level architectural principles where models of self-organizing topographically ordered representations are building blocks. Since I adhere to a systems-level modeling approach, subsystems of the cognitive architecture can be updated and substituted in an iterative fashion for improvement.

## 3. HIERARCHICAL FEATURE REPRESENTATIONS

How can self-organized topographically-ordered representations of a more abstract kind, e.g. a representation with semantically related symbols that occupy neighboring places be obtained in a cognitive architecture?

In the mammal brain there seems to be a principle of hierarchical ordering of representations, e.g. the executive and motor areas seem to be hierarchically ordered from more abstract to less abstract representations. Constraining the discussion to the perceptual parts of the mammal brain, we find that the different sensory modalities (visual, somatosensory, auditory, ...) adhere to a hierarchical organizational principle. For example, we find hierarchically-organized topology and probability-density preserving feature maps in the ventral visual stream of the visual system. These feature maps rely on the consecutive input from each other and tend to be hierarchically-ordered from representations of features of a lower complexity to representations of features of a higher complexity. Thus, we find ordered representations of contour directions in V1 in the Occipital lobe, of shapes in V2, of objects in V4, and of faces or complex facial features in the inferior temporal (IT) area of the Temporal lobe.

The hierarchical organization principle is employed artificially in Deep Neural Networks, i.e. in artificial neural networks with several hidden layers. A neural network that has been applied very successfully within the field of computer vision is the Deep Convolutional Neural Network (LeCun et al., 1998).

Here, when I discuss the hierarchical ordering principle for perceptual parts of a bio-inspired cognitive architecture, this principle is instantiated by hierarchical SOMs. The choice of SOMs is not based on performance, but on the fact that the hierarchical organization principle is also to be combined with other principles in the cognitive architecture elaborated on below. For the moment the SOM and its variants are considered good choices to explain and test principles.

Together with collaborators, the author has shown the validity of this hierarchical organizational principle repeatedly with hierarchical SOMs when applied to different sensory modalities. For example, in the case of the somatosensory modality, several experiments have been conducted to show how haptic features of an increasing complexity can be extracted in hierarchical self-organizing representations, e.g. from proprioceptive and tactile representations at the lower complexity end to self-organizing representations of shapes and sizes of the haptically explored objects (Johnsson et al., 2011a). Another example in the case of the visual domain where experiments have been done to show that hierarchies of ordered representations of postures at the lower complexity end to ordered representations of gesture movements of the observed agent can be self-organized (Buonamente et al., 2016).

## 4. SUPPLEMENTING SENSORY SIGNALS IN PERCEPTION

The brain supplements perceptions when the sensory input is not complete. This is evident from various visual illusions, e.g. the Kanizsa Triangle (Kanizsa, 1976) where the contours of a triangle can be perceived even though they are actually not there. Moreover, when our eyes scan the scenery before us, they are doing so by semi-random eye movements known as saccades directing the movements toward particularly conspicuous and, in some sense interesting features. Supposedly we carry out similar semi-random movements with our hands and fingers to gain particularly interesting and useful tactile sensory input when we, for example, ransack our pockets for a particular key, or grope about to find the doorknob in the dark. When we perceive our brains seem to fill in the gaps of sensory input with expectations, from memory, of what is likely to be there.

A crucial aspect of biological cognition is an ability to simulate or influence perceptual activity in some brain areas due to the activity in other brain areas (Hesslow, 2002; Grush, 2004), e.g. the activity in areas of other sensory modalities. For example, when the visual perception of a lightning evokes an expectation of the sound of thunder, or when visual images/expectations of an object is evoked when it is felt in the pocket. Hence, one supplement to the afferent sensory signals in perception could be such simulated *Cross Modal Expectations*. These could even override actual input, which is evident from the McGurk-MacDonald effect (McGurk, MacDonald, 1976). If a person sees a video with someone making the sound /da/ on which the lips cannot be seen closing and the actual sound played is /ba/, the expectations evoked by the visual perception may have such an influence on the activity caused by the actual afferent auditory sensor signals that the person may still hear the sound /da/.

A variant of the SOM, the A-SOM (Johnsson et al., 2009), that adds adaptable associative connections between feature representations has been used to build artificial systems, e.g. (Johnsson, Balkenius, 2008), that demonstrate the supplementation of sensory input and the elicitation of cross-modal expectations.

## 5. NETWORKS OF FEATURE REPRESENTATIONS

In perception, sensory signals from receptors, together with information about involved exploratory actions, such as eye or hand movements, activate sets of feature maps. Those parts of the associated networks of feature representations that are not elicited directly by sensory input, are activated through the activity in other feature representations via associative connections. Hence the perceptions will be complete even with scarce sensory in-

put, because missing parts are filled in with likely guesses through internal simulations.

Through adaptable associative connections between hierarchies of topographically ordered feature representations self-organizing intra- and intermodal *Networks of Feature Representations* (NFRs) are obtained. Some feature representations can be part of several NFRs, and the particular division of the feature maps into NFRs depend on how we look at it and how we choose to categorize the system into subsystems. The adaptive associative connections learn to associate simultaneous, or temporarily close, activity in various feature representations elicited by simultaneous, or temporarily close, but different ordinary input. This means that feature representations that later lack ordinary input will be activated by activity patterns associated with the ongoing activity in other feature representations in the NFR. For example, hearing the voice of a particular person would elicit activity patterns not only in the auditory hierarchies of feature representations that directly receives sensory input, but also in other, e.g. visual, feature representations in an intermodal NFR through associative activation. The total activity in the NFR will constitute episodic memories, imagination etc.

## 6. MEMORY

Although there are hierarchically-organized feature representations in the brain, it is questionable whether there are neurons—aka grandmother cells—that are the exclusive representatives of distinct individual objects. Though there is no total consensus regarding this, I consider it more likely that distinct individual objects are coded in a more distributed way as an ensemble of feature representations in the NFR, at various complexity levels, across several sensory (as well as non-sensory) modalities. Hence, the recognition of distinct individual objects consists in the simultaneous activation of a sufficiently large and unique subset of this ensemble of representations across various modalities.

Thus, the representation of a real or imagined concept or object is constituted by a set of associated activity patterns in various feature representations of the NFR distributed over multiple modalities. Such associated activations of topologically ordered feature representations preserve an internal ordering of activation and could be seen as forming a *Conceptual Space* (Gärdenfors, 2000).

The activation of some feature representations will tend to trigger expectations/imaginations of features of the distinct individual object in other representations across various modalities, presumably associated by associative connections in a way similar to the activation of more features of high-

er—or lower—complexity in hierarchically connected feature representations (which can as well be cross-modal).

I believe that such connectivity—associative and hierarchical—between feature representations of various modalities and at various complexity levels are what enables the filling in of missing parts of our perception by imagination, but also that they enable our various kinds of memory.

Different kinds of memory are, I believe, using the same kind of feature representations across modalities in our brains. What differs between different kinds of memory is rather how they are activated and precisely what ensemble of feature representations in the NFRs that are activated.

For example, one could speculate—in a simplified way—that the working memory supposedly again employs networks of the same building blocks of feature representations obtained during early developmental phases, but now activated in a more transient and temporary way, perhaps from, in the case of the mammal brain, the frontal lobes, whereas perception as such, is the activation of an ensemble of feature representations due to afferent sensory signals, together with the filling-in of missing parts due to cross-modal as well as top-down expectations at various levels of hierarchies.

In episodic memory and imagination (i.e. internal simulation) the sets of associated networks of feature representations (which can also be non-sensory, such as motor representations) are activated internally (at least partly) in the cognitive architecture/brain. The associatively connected representations (actually associated activity patterns in the underlying wetware / hardware) are what lends memory and imaginations their associative characters.

The important point here is that it is reasonable to believe that the same (simple) principles are behind both the supplementation of perception and the associative character of memory and imagination; and the distributed, associative and hierarchical character of the intra- (and inter) modal representations they all (perception, imagination and memory) rely on. That different faculties go into each other also explains why we tend to both keep memories alive (by strengthening associative connections through reactivation) and sometimes change them over time (via imagination) when we recapitulate.

Semantic memory presumably corresponds to more persistent associations due to repeatedly overlapping activity from many various perceptual and episodic examples over time, thus forming prototypes in conceptual spaces. This also makes semantic memory more persistent, as well as more resistant in a deteriorating/aging system.

The point here is that all kinds of memory, perceptions, imaginations, and expectations are proposedly using simultaneous and / or sequential activations of ensembles / subsets of the same NFRs across various modalities in the brain. In fact, I think that there is no reason that the representa-

tions should be constrained to the brain only, but that associated representations could also be various kinds of activity in / of the body, such as e.g. postural / breathing patterns, hormonal configurations etc. This would also explain why the change of posture / breathing patterns can change the state of the mind. In the extension, even "representations" that we interact with—and continuously reconfigure—in the environment outside the body—including the representations within other agents, such as humans, pets, machines etc.—are presumably included.

To learn to represent novel concepts, objects or possible objects, there is no need for new feature representations, because they are formed through associating activity patterns in existing feature maps in novel ways.

This kind of feature ensemble coding also enables / explains the ability to represent completely novel categories / concepts in the brain / cognitive architecture, and the ability to create and imagine non-existing and perhaps impossible concepts, objects, etc. This is because representations are composed of sufficiently large ensembles of associated multi-modal features, and novel associated ensembles and sometimes associated ensembles corresponding to concepts and imaginations that do exist (but have not yet been seen or reached) or do not exist in our physical reality (e.g. unicorns) can emerge.

Of course, there are limits to what we can imagine and conceptualize, and perhaps even think about. For example, we are unable to visualize objects in spaces of a higher dimensionality than three. However, such limitations are just to be expected if all perceptions, memories and imaginations are made up of distributed (in space and time) activations of ensembles of associated features, and there are constraints on what kind of features can be represented in the brain (or cognitive architecture), which is likely. The constraints are probably set by biological limitations that exist due to a lack of evolutionary pressure, as well as determined by the development of the organism in its environment. An example of the latter is that cats raised in an environment consisting entirely of vertical lines during a critical developmental phase during infancy will be unable to see horizontal lines (Blakemore and Cooper, 1970). That there are constraints on what kind of features that can be represented also implies the possibility that all that we can think about regarding reality is not necessarily corresponding to all that there would have been to think about, had we been wired differently.

In accordance with the reasoning above, it is reasonable to assume that the need for associative connections—corresponding to axon bundles in the neural system of a biological organism—between feature maps at various complexity levels within as well as between different modalities are of significance in a cognitive architecture based on self-organizing topographically-ordered feature representations. Such associative connections need to be adaptive (by adjustable parameters corresponding to modifiable synapses in

the neural system of a biological organism) to enable the learning of associations between the activity in various feature representations.

## 7. IMAGINATION

In addition to an ability to automatically develop, and continuously readapt, sensory and other representations, and their interconnections that connect simultaneous activity within them spatially, a bio-inspired autonomous agent needs an ability to learn to associate activations of representations over time. This is desirable because it enables the autonomous agent to remember and re-enact sequences of perceptual—and other—activity across modalities and levels of hierarchy.

With such an ability an autonomous agent can remember sequences of perceptions, and if the ability is generalized, other things as well, e.g. motor activities. Such perceptual sequences could, for example, be visual landmarks. To the perceived visual landmarks, appropriate motor activity could be associated. With perceptual sequences simultaneously learned in other modalities together with cross-modal associations, the sequential memories are reinforced and thus diminish the influence of noise and limitations in sensory input. The perceptions (and preparatory responses etc.) corresponding to missing input in some modalities—sensory and other—will be imagined, i.e. elicited through cross-modal activation. If suddenly the agent would lack input to some, or all, sensory modalities, it would still be able to operate and to some extent carry out actions associated with imagined perceptions of the environment. With this kind of ability an agent would also be able to sit idle imagining various scenarios and the likely consequences of carrying out different kinds of actions. The latter is valuable for survival and will also accelerate the agent's learning.

The idea to internally elicit activity patterns in perceptual, motor and other circuits over time (activation sequences) and in space (in various feature maps across different modalities), corresponding to the patterns that would have been elicited had there been sensory input and had the actions been carried out, is closely related to the simulation hypothesis by Hesslow (2002). It could in the extension also be the foundation for providing agents with an ability to guess the intentions of other agents, either by directly simulating the likely perceptual continuations of the perceived behavior of an observed agent, or by internally simulating its own likely behavior in the same situation under the assumption that the other agent is similar in its assessments, experiences and values that drives it.

A mechanism that implements self-organizing topographically ordered feature representations that can be associatively and recurrently connected with an arbitrary number of other representations and with arbitrary time

delays is the Associative Self-Organizing Map (A-SOM). Hence the A-SOM would in some cases be a better choice, than the standard SOM, to use as one of the basic building blocks in the perceptual parts of a cognitive architecture. An A-SOM can learn to associate the activity in its self-organized representation of input data with arbitrarily many sets of parallel inputs and with arbitrarily long-time delays. For example, it can learn to associate its activity with the activity of other self-organization maps, or with its own activity at one or more earlier times. This allows for cross-modal expectations. For example, if a sensory modality, say the visual system in a cognitive architecture, produces a certain internal pattern of activity due to sensory input, then activity patterns are elicited in other sensory modalities corresponding to the patterns of activity that are often triggered in these other sensory modalities through sensory inputs that usually occur simultaneously, even when they do not. Due to the ability of the A-SOM to associate its activity with its own activity at one or more earlier times, a mechanism for sequence completion that can be used for internal simulation is made possible. This is consistent with those abilities necessary for an autonomous agent described above. The A-SOM has been successfully tested in many simulations (e.g., Johnsson et al., 2011b) in several different domains, as well as together with real sensors such as tactile sensors (Johnsson, Balkenius, 2008) and cameras (Buonamente et al., 2015), and when simulating likely continuations of sequences of strings of symbols and words (Gil et al., 2014). It has been used to simulate the sensory activity patterns likely to follow some initially perceived movements of actions/gestures (Buonamente et al., 2015). In the domain of music, a further developed and more mature and generalized version of the A-SOM has been used to simulate the sensory activity patterns likely to follow those elicited by the initial parts of perceived Bach chorale melodies (Buonamente et al., 2018).

Associative connections are in place between different representations at various levels of feature complexity. Simultaneously-activated feature representations develop stronger associative connectivity. The result is that we will find strongly interconnected sets of feature representations—and other kinds of circuits—in the brain/architecture. As humans, we label these and call them systems/components of one kind or another (depending on the particular discipline, the prevailing paradigm and zeitgeist), though we should keep in mind that these categorizations and demarcations are our inventions and thus somewhat arbitrary.

The inter-connectivity of the feature representations within a modality/ submodality tend to be strong because it has been reinforced by simultaneous activations originating from the receptors of the modality specific sensory organs. Thus, connective configurations / subsystems in the brain / architecture develop through the repeated simultaneous activation of sets of self-organizing feature representations.

However, the feature representations within a modality also connect to feature representations in other modalities / systems, only to a lesser extent. This is due to the statistically fewer simultaneous activations of feature representations in other modalities. Various systems activate each other through these associative connections that have learned to associate activity that normally come together. Hence, if the activity within one system, perhaps triggered through afferent signals from sensory organs or from some other part of the brain/architecture, tend to correlate with the activity of other systems, perhaps triggered by the afferent signals from other sensory organs or other parts of the brain/architecture, then the inter-connectivity of the systems is reinforced. The foundation for these correlated activities in various systems is that sensory stimuli, and the consequences of an agent's actions, are related in a non-random way due to the statistical regularities of the properties of the world. These statistical regularities will be reflected in the associative connectivity between various systems.

In reality the various cognitive functions are not separated from each other in a neat way. Rather, they blend and mix into each other. For example, the perception of hearing a familiar person's voice can trigger both episodic memories, internal visual simulations of the person, corresponding to reality but also pure fantasies, etc. Internally simulated perceptual expectations in turn may trigger exploratory behavior and attention aiming at confirming the expectations by obtaining additional sensory input. All this founded on associatively connected networks of topologically ordered feature representations.

Taken together, all this means that NFRs containing topologically ordered feature representations with intra- and intermodal adaptable associative connections enable perception, various forms of memory and imagination. In addition, it provides a mechanism for representing the ongoing activity in one system/NFR with the activity of other systems/NFRs.

## 8. CONSCIOUSNESS

Consciousness is about experiencing perceptions, including the perceptions of our own actions; imagery; memories. But who is experiencing it? I am considering functional consciousness here, thus leave the problem of qualia out of the discussion.

In the discussion about cross-modal expectations and internal simulations above, I discussed how activity in some feature representations can elicit reasonable activity in other feature representations through associative connections. The elicited activity in the latter representations correspond to the activity that normally would or could occur simultaneously, or timed,

with the activity in the first representations even though the latter lack any afferent input ultimately originating from sensors.

I believe that the same mechanism with adaptive associative connections in the case of a bio-inspired cognitive architecture, or nerve bundles with synapses in the case of a neural system of a biological organism, between different subsets of feature representations, at various levels of abstraction, is significant for the realization of at least p-consciousness. From this perspective, the elicitations of activity in some feature representations by the activity in other feature representations via associative connections can be viewed as if the activity in the latter system (composed of the activity in connected, perhaps distributed, feature representations), in a sense, is represented by the activity in the former system (of associatively connected feature maps).

Various systems could perhaps also 'observe' each other simultaneously as well. The mechanisms and principles sketched above could be behind or be used for a kind of summarization of the observed subsystem's or subsystems' activity at a possibly different and more abstract level.

As also argued by Hesslow and Jirenhed (2007), perceptual simulation could explain the appearance of an inner world. A remaining question is 'who' is observing regardless of whether it is perceptions ultimately elicited from sensory organs, internal simulations originating from within the brain, or some combination thereof. My proposal is that they are observed by other connected configurations of systems whose activity summarizes/represents the observed internal simulations and perceptions, because their corresponding activity correlates due to the learning represented in the adaptive associated connections. The same systems could perhaps have multiple functions while also "observing" each other simultaneously as well.

Put differently, this could be seen as one system observing the various *Phenomenal Maps* of another system, whether these are activated due to sensory signals or through internal simulations (imagination, episodic memory, working memory etc.).

Still another way to put it is that some systems are aware of, i.e. p-conscious / functionally conscious of, other systems' perceptual activity.

The activity of associatively connected configurations of feature representations correlates because the adaptations of the associative connections between the representations, and the adaptions of the representations themselves happens simultaneously, continuously and dynamically. At a lower perceptual level this means that the activation of feature representations in some sensory modalities will elicit activity in feature representations in other sensory modalities and consequently sensory expectations / supplementations in those other modalities, as discussed above.

Thus, I believe that adaptive associative connections between and within various configurations of strongly connected feature representations at vari-

ous levels of complexity or abstraction are of significant importance for realizing p-consciousness, i.e. functional consciousness, in a cognitive architecture.

## REFERENCES


C. Blakemore, G. F. Cooper, *Development of the Brain Depends on the Visual Environment*, Nature, 228, 1970, pp. 477–478.

N. Block, *On a Confusion about a Function of Consciousness*, Behavioral and Brain Sciences, 18, 1995, pp. 227–287.

P. Bołtuć, *Strong Semantic Computing*, Procedia Computer Science, 123, 2018, pp. 98–103.

____ , *The Philosophical Issue in Machine Consciousness*, International Journal of Machine Consciousness, 1 (1), 2009, pp. 155–176.

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

M. Buonamente, H. Dindo, A. Chella, M. Johnsson, *Simulating Music with Associative Self-Organizing Maps*, Journal of Biologically Inspired Cognitive Architectures, 25, 2018, pp. 135–140.

M. Buonamente, H. Dindo, M. Johnsson, *Discriminating and Simulating Actions with the Associative Self-Organizing Map*, Connection Science, 2 (27), 2015, pp. 118–136.

M. Buonamente, H. Dindo, M. Johnsson, *Hierarchies of Self-Organizing Maps for Action Recognition*, Cognitive Systems Research, (39), 2016, pp. 33–41.

D. J. Chalmers, *Consciousness and Its Place in Nature*, in: Blackwell Guide to the Philosophy of Mind, S. Stich, T. Warfield (eds.), Blackwell Publishing, Malden, MA 2003, pp. 102–142.

A. Damasio, *Self Comes to Mind: Constructing the Conscious Brain*, Pantheon, Cambridge 2010.

D. van Essen, *Functional Organization of Primate Visual Cortex*, Cerebral cortex, 3, 1985, 259–329.

B. Fritzke, *Growing Cell Structures—A self-organizing Network for Unsupervised and Supervised Learning*, Neural Networks, 9 (7), 1994, pp. 1441–1460.

____ , *Growing Grid—a Self-organizing Network with Constant Neighborhood Range and Adaptation Strength*, Neural Processing Letters, 5 (2), 1995, pp. 9–13.

P. Gärdenfors, *Conceptual Spaces—The Geometry of Thought*, MIT Press, Cambridge, MA 2000.

D. Gil, J. Garcia, M. Cazorla, M. Johnsson, *SARASOM – A Supervised Architecture based on the Recurrent Associative SOM*, Neural Computing and Applications, 5 (26), 2014, pp. 1103–1115.

R. Grush, *The Emulation Theory of Representation: Motor Control, Imagery and Perception*, Behav. Brain. Sci., 27, 2004, pp. 377–442.

G. Hesslow, *Consious Thought as Simulation of Behaviour and Perception*, Trends Cogn. Sci., 6, 2002, pp. 242–247.

G. Hesslow, D.-A. Jirenhed, *The Inner World of a Simple Robot*, J. Consc. Stud., 14, 2007, pp. 85–96.

M. Johnsson, C. Balkenius, *A Robot Hand with T-MPSOM Neural Networks in a Model of the Human Haptic System*, in: the Proceedings of Towards Autonomous Robotic Systems 2006, 2006, pp. 80–87.

____ , *Sense of Touch in Robots with Self-Organizing Maps*, IEEE Transactions on Robotics, 3 (27), 2011a, pp. 498–507.

____ , *Associating SOM Representations of Haptic Submodalities*, in: The Proceedings of Towards Autonomous Robotic Systems 2008, 2008, pp. 124–129.

M. Johnsson, C. Balkenius, G. Hesslow, *Associative Self-Organizing Map*, in: The Proceedings of the International Joint Conference on Computational Intelligence (IJCCI) 2009, 2009, pp. 363–370.

M. Johnsson, M. Martinsson, D. Gil, G. Hesslow, *Associative Self-Organizing Map*, in: Self-Organizing Maps—Applications and Novel Algorithm Design, MA: Intech, 2011b, pp. 603–626.

G. Kanizsa, *Subjective Contours*, Scien. Am., 234 (4), 1976, pp. 48–52.

T. Kohonen, *Self-Organization and Associative Memory*, Springer Verlag, Berlin–Heidelberg 1988.

Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *Gradient-Based Learning Applied to Document Recognition*, in: Proceedings of the IEEE, 1998, pp. 2278–2324.

H. McGurk, J. MacDonald, *Hearing Lips and Seeing Voices*, Nature, 264, 1976, pp. 746–748.

R. A. Reale, T. H. Imig, *Tonotopic Organization in Auditory Cortex of the Cat*, J. Comp. Neurol., 192, 1980, pp. 265–291.

H. Ritter, T. Kohonen, *Self-Organizing Semantic Maps*, Biol. Cybern., 61, 1989, pp. 241–254.

A. R. Tunturi, *Physiological Determination of the Arrangement of the Afferent Connections to the Middle Ectosylvian Auditory Area in the Dog*, Am. J. Physiol., 162, 1950, pp. 489–502.

A. R. Tunturi, *The Auditory Cortex of the Dog*, Am. J. Physiol., 168, 1952, pp. 712–717.

S. Zeki, *The Representation of Colours in the Cerebral Cortex*, Nature, 284, 1980, pp. 412–418.

ABOUT THE AUTHOR — a cross-disciplinary researcher. He is currently Associate Professor in autonomous systems at Malmö University in Sweden and Researcher at Magnus Johnsson AI Research AB. More information at www.magnusjohnsson.se

Email: magnus@magnusjohnsson.se

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Rafał Maciąg

# TOWARDS THE PRAGMATIC CONCEPT OF KNOWLEDGES

doi: 10.37240/FiN.2022.10.zs.11

***ABSTRACT***

The article presents and justifies the thesis that the way of understanding knowledge has changed significantly over the last century. This change consists in departing from the classic definition of knowledge formulated by Plato, and in particular in questioning the subjective role of man as the holder of knowledge and abandoning claims to the truthfulness of knowledge. This process was an intensive evolution; its elements are given and justified in the text. Its source was a deep reconstruction of the mode of creating epistemic structures in mathematics and geometry, based on the abandonment of the principle of representation. Knowledge turned out to be determined by the social context, it became dispersed, decentralized, which led to the rejection of the condition of its truthfulness. The last phase of this evolution is knowledge as a phenomenon in the area of digital technologies, in particular artificial intelligence. This evolution has led to the emergence of many variants of knowledge that act as local knowledge, which justifies the use of the plural in this case.

**Keywords**: knowledge, metamathematics, artificial intelligence, sociology of knowledge, truth.

## 1. INTRODUCTION

This text attempts to justify an observation about the rapid change in the approach to the phenomenon of knowledge in the twentieth century; the change leads to the emergence of various variants of understanding knowledge, founding various social phenomena, and thus justifying the need to use the term “knowledges” (plural ). Most of this movement takes place as a certain historical and social process, which draws attention to the pragmatic nature of the transformations and mainly concerns propositional knowledge, but, of course, it must also affect theoretical reflection. After all,

knowledge about knowledge is also a kind of knowledge, so the point of view based on the opposition between man and the world cannot be omitted here. It is fundamental to philosophy, as a prototype situation from the point of view of knowledge, interpreted as an effort to get to know the world by man. This kind of observation opens up an extremely extensive problem, exceeding the scope of this study, which focuses on gathering the basic premises of the observation made and putting in its light the thesis about the transformation of the way of understanding knowledge. This transformation at the starting point is based on the perception of knowledge as a certain state, resulting from communing with the world. Such thinking has its origins in the Platonic concept and persists until the end of the nineteenth century. The present state of the transformation mentioned above brings the idea of dispersed, diversified, and decentralized knowledge. The traditional quality of knowledge, which is its truth, is necessarily eroded. The concept of knowledge is also becoming the subject of intense analytical effort, which is reported, among others, by Burgin (Burgin, 2015).

In light of the emerging analyzes and phenomena, the issue of articulating knowledge, as well as its retention, has become significantly complicated. One of the most important contemporary fields in which the pragmatic approach to knowledge has led to unusual and previously unheard of implementations is technology, or more precisely digital technology. Philosophical attempts to combine the issues of knowledge and mathematics or logic are old and date back at least to the end of the nineteenth century, and taking into account the visions of Leibniz and Descartes, even earlier (Russell, Norvig, 2010). However, direct, technical attempts to formalize knowledge within computational technology appeared at the beginning of the second half of the twentieth century, opening a process that was revealed in an advanced way in the case of the so-called artificial intelligence. Due to the spectacular development of this field, especially within the so-called natural language processing, the topic of knowledge formalization seems to be one of the most important contemporary contexts of knowledge issues.

Knowledge has been the subject of philosophy since the beginning of European culture. Although a separate branch of philosophy called epistemology, understood as a self-conscious research procedure, is much younger, the concept of epistemology appeared in the 18th century. The most important questions that arise within it are given by Wolenski: “What is knowledge?; Is knowledge based on senses or reason? Is certainty attainable? What is truth? Are there ultimate limits of knowledge?” (Woleński, 2004, p. 4). Chisholm asks analogous questions somewhat differently, directly referring to the Platonic definition: “What can I know? How can I distinguish those things I am justified in believing from those things I am not justified in believing? And how can I decide whether I am

more justified in believing one thing than in believing another?" (Chisholm, 1989, p. 1).

Chisholm shows a clear shift of the problem towards the subject, which is indicated as the central instance. A similar view seems to prevail, especially in the basic or introductory approach, i.e. where we are not dealing so much with theory but rather with reporting the current, valid explanation. In a textbook on epistemology, Robert Audi expresses the underlying assumption directly. Giving examples of sources of knowledge, he lists the first and most important ones: perception, and then others: "memory as a storehouse of what we have learned in the past, consciousness as revealing our inner lives, reflection as a way to acquire knowledge of abstract matters, and testimony as a source of knowledge originally acquired by other people" (Audi, 2003, p. 1). By giving examples, he immediately refers to the human being. Writing how knowledge is produced, he emphasizes three acts in the course of which knowledge is perceived, believed, and justified by a certain ego. Duncan Pritchard confirms a strikingly similar approach: "Two things that just about every epistemologist agrees on are that a prerequisite for possessing knowledge is that one has a belief in the relevant proposition and that that belief must be true" (Pritchard, 2006, p. 5).

The views just quoted have one source, which is Plato. His views on knowledge are scattered throughout many dialogues, such as Phaedo, Symposium, Republic, Timaeus, Sophist, or Statesman (Burgin, 2015; Cornford, 1935). However, the best-known and quoted definition of it comes from Theaetetus, who expresses it in a dialogue with Socrates as overheard from someone whom the hero did not remember. In literature, it appears as a short phrase: *Dóksa alethés metá lógu*. Its English version, shortened to three words, reads *Justified True Belief*. It was mentioned here because Appiah argues that it "is a central philosophical claim [about knowledge] of the Western tradition since Plato" (Appiah, 2003, p. 43). The Polish translation is provided by Władysław Witwicki: "wiedza to jest sąd prawdziwy, ściśle ujęty" (Platon, 2002, p. 178). At the same time, he describes in a footnote the translation problems provided by the last term in the Greek version, derived from the term "logos" (*λόγος*). The choice made by Witwicki, i.e. this "ścisłe ujęcie," corresponds with other traditions in this regard, e.g. the German: *das Wesen der Erklärung* (Plato, 1856, p. 208), English: *rational account* (Plato, 1987, p. 115), or French: *l'essence de leur definitione* (Deschoux, 1980, p. 289). Each time it is about a certain technical skill related to order and discipline of reasoning. For Plato, knowledge is a purely human disposition, and its status remained unchanged for over two thousand years. This assumption is emphasized by the use of the term *dóksa*. However, this quality of knowledge, as well as its second quality, i.e. truthfulness (*alethés*), will be challenged in the process of reconstructing knowledge understanding that took place in the 20th century.

In the twentieth century, the concept of knowledge began to be applied in new and hitherto unprecedented circumstances, in areas such as politics, the management, or the broader social life, taking on a slightly different shade in each of these contexts. It seems that the reason for this was undermining the key epistemic competence of man, which is the ability to express justified and true beliefs about the world, entering the area of no less basic and intuitive philosophical construction, which is a man confronted with the world. Such thinking is as old as Western culture, and its dominant rationalist interpretation is given by Descartes, who formulates the dualism of matter (*res extensa*) and mind (*res cogitans*) (Descartes, 1641). In this way, the mutual positioning of man and the world for the next 250 years is determined, creating a context for understanding knowledge that is the property of the mind and at the same time presupposes the existence of the world as a certain objective entity that can be comprehended by reason.

Around the second half of the nineteenth century, such an assumption rapidly eroded, caused by an epistemological crisis within mathematics and geometry, leading to a profound, paradigmatic shift in the way concepts that interpreted the nature of the world were constructed. This "digital transformation" (Maciąg, 2020), with deep epistemological foundations, also becomes an involuntary source of technology in which knowledge, due to practical necessity, is formalized, allowing for calculations to be made with the use of this knowledge. This understanding of knowledge is completely detached from the human being as its user and disposer, replaced by a more or less advanced artificial cognitive system. Knowledge in such a process is instrumentalized and adapted to the purposeful and utilitarian requirements of computing systems, performing clearly defined tasks of an "intelligent" nature, such as translating a text, answering questions about its content, understanding it, supplementing it, etc.

The next step, which also tears the notion of knowledge from the subject, is to place knowledge in a broader social context, which is formulated at the beginning of the twentieth century and developed in various ways later. This kind of conditioning has resulted in a multiple and dispersed image of knowledge, strongly relativized to the circumstances of its use. At the same time, this relativization resulted in a departure from the requirement of the truthfulness of knowledge, understood as the ability to represent the world. In its place, local and *ad hoc* variants appeared, causing the multiplication of many, numerous, and contingent knowledge in place of one true knowledge, producing a state of "the differend" (*le différend*), referring to the nomenclature given by Lyotard (Lyotard, 1983).

The interpretation proposed in this text, according to which knowledge is perceived as dispersed and multivariant while losing the connection with a man as its disposer and the attribute of truth in the sense of representing the world, is long-lasting and lasts from the end of the 19th century to the

end of the 20th century, but it brings about profound and significant effects. This interpretation is based on a pragmatic approach, based on an idea contained in the writings of philosophers such as John Dewey and Charles Sanders Peirce. Dewey writes: "In order to be able to attribute a meaning to concepts, one must be able to applay them to the existence. Now it is by means of action is made possible" (Dewey, 1931, p. 15). Peirce thinks similarly, writing a little earlier: "the rational purport of a word or other expression, lies exclusively in its conceivable bearing upon the conduct of life" (Peirce, 1905, p. 162). The study of the way of understanding knowledge based on a pragmatic approach necessarily becomes concluding the way of applying this concept, engaging social and historical contexts.

## MAIN PRINCIPLES FOR A CHANGE IN THE WAY OF UNDERSTANDING KNOWLEDGE

The described process of changing the way of understanding knowledge and thus giving meaning to its concept, consisting in observing the ways of its use, as well as the effects it causes, is long and complex. The way to solve this problem is to present the main areas in which this change takes place. They provide specific premises for the proposed interpretation of the phenomenon of knowledge, although due to the breadth of this process, they are of various nature. They should be treated as important milestones in transformation of the understanding and use of the concept of knowledge, which, following the concept adopted here, are forming a coherent trend. The aforementioned premises are based on the following facts: (1) the epistemological paradigm shift that took place in the fields of mathematics and geometry in the second half of the 19th century and the beginning of the 20th century; (2) Karl Popper's concept which practically excludes the possibility of expressing constructive and sure judgments about the world; (3) The so-called Gettier problem, triggered by his article in which he shows the inconsistency of the Platonic definition of knowledge (Gettier, 1963); (4) a great project of relativizing human epistemic abilities and knowledge in relation to the social context, having many variants, developed in the 20th century; (5) pragmatic understanding of knowledge as a resource that can shape large-scale social phenomena, such as the knowledge society, perform the function of an organizational resource in economic processes, which has been collected under the banner of Knowledge Management (KM) or generate arbitrary, practical orders of knowledge what the field of Knowledge Organization (KO) does; (6) the emergence of knowledge as a resource that is the basis of artificial cognitive systems, with various levels of advancement, also as developed as the latest language models in the area of natural language processing (NLP).

1. First of all, the notion of change in the epistemological paradigm that took place at the end of the 19th century and the beginning of the 20th century due to mathematics and geometry needs to be clarified. Morris Kline places the deep causes of this change in the emerging ideas of the so-called non-Euclidean geometries. The two main projects of them by Nikolai Lobachevsky and János Bolayi appeared in the first half of the nineteenth century although they are the result of uncertainty about the so-called fifth postulate of Euclid's geometry, which appeared much earlier (Murawski, 2001). Their basic and simplest property was to completely ignore the experience of the geometry of the world, which was then apparent to the observer, and today it defines the common experience in this regard. These ideas opened the way to thinking in terms detached from this experience and opened up the possibility of reasoning unlimited by the necessity of conforming to it. Kline describes it as follows: "The two-thousand-year-old conviction that mathematics was the truth about nature was shattered. But the mathematical theories now recognized to be arbitrary had nevertheless proved useful in the study of nature. Though existing theories historically owe much to suggestions from nature, perhaps new theories constructed solely by the mind might also prove useful in the representation of nature. Mathematicians then should feel free to create arbitrary structures" (Kline, 1990, p. 1036).

Similar fundamental effects are noted by Luke Hodgkin in the context of the concept of numbers appearing in the works of Richard Dedekind, Gottlob Frege, and Giuseppe Pean. He wrote that the situation "did lead to a reshaping of mathematics if not the whole world-view" (Hodgkin, 2005, p. 215), because they have resulted in a "crisis of foundations," and thus the meaning and justification of basic material beings. They lost their ontological basis and became only assumed theoretical constructions: "the objects of mathematics were not actual things-in-themselves (as one thinks of a triangle, say, or the number '7'), but the rules which they obeyed" (Hodgkin, 2005, p. 216). From now on, mathematics and geometry cease to be stories about the world, they cease also to represent it, but when freed, they can produce their own structures, based solely on internal coherence resulting from the adopted assumptions. Kline describes the situation as a "loss of truth" and concludes that "By 1900 mathematics had broken away from reality; it had clearly an irretrievably lost its claim to the truth about nature, and had become the pursuit of necessary consequences of arbitrary axioms about meaningless things" (Kline, 1990, p. 1035). Although the very concept of the axiom is derived from Proclus, the independent authors of their contemporary idea are Giuseppe Peano (Peano, 1889) and David Hilbert (Hilbert, 1899).

The strategy of free, only internally disciplined speculation turns out to be effective. Not only does it not prevent their application to the physical

world, but it also allows one to overcome the disadvantages of human observation resulting from imperfect tools of cognition. The reversed direction of building models of the world necessarily has distant and, so to say, opposite effects: it causes a deep crisis of trust in various forms of apperception. This apperception covers not only the direct world of entities but also the rules governing that world, including the cause-and-effect principle as the basis for the functioning of the world. This fundamental breakthrough concerning human cognitive competencies and sources of knowledge makes its way into other areas of reflection, also of a humanistic nature, resulting in the so-called poststructuralism and postmodernism. These kinds of thread appear in the writings of Jean-François Lyotard, Jeaques Derrida, as well as Gilles Deleuze, and Michel Foucault. The questioning of the cause-and-effect principle is realized directly as a completely new idea of complexity, which is built on two new descriptive structures that appeared at the beginning of the 20th century: network theory and systems theory. The idea of complexity allows one to understand the processes and phenomena of very different areas of reality, proposing an interpretation with a very high degree of universality.

The fruit of the mathematical debates that emerged at the end of the nineteenth century was, *inter alia*, Hilbert's program, consisting of 23 points, which, as Murawski writes, was "an attempt to justify the classical (infinite) mathematics and to save its integrity by showing that it is secure" (Murawski, 2010, p. 29). It aimed to stabilize and justify mathematical procedures and constructions by introducing defined axioms and proving rules, allowing one to build the foundations of mathematics. In section ten, Hilbert formulates the question which Roger Penrose summarizes as follows: "is there some general mechanical procedure which could, *in principle*, solve all the problems of mathematics (belonging to some suitably well-defined class) one after the other?" (Penrose, 1999, p. 34). This question was taken up by Alan Mathison Turing. To prove the impossibility of such a procedure, he proposes a theoretical machine that becomes a conceptual prototype of a modern computer. Its computational mode of operation determines its further development, which also applies to knowledge, which is subject to appropriate reconstruction (formalization) to meet technical requirements. The invention of the computer and the subsequent incorporation of the phenomenon of knowledge into digital technology, therefore has the same origins as the fundamental and less known epistemological revolution that took place in mathematics and geometry. It also implements an analogous set of epistemological assumptions, based on arbitrary interpretative procedures, which at the epistemic level are conceptual in nature, but at the level of the technical apparatus, they come down to certain practical procedures (the so-called digitization). Both variants strongly influence the phenomenon of knowledge, leading to its actual instrumentalization, taking place at different levels of interpretation: theoretical and practical.

2. The described revolution becomes the source of a change in the approach to the place which is by definition the field of creating knowledge, i.e. science, which in this text represents the second premise. In this way, one can interpret the concept of Karl Popper, which he presented in *Logik der Forschung: Zur Erkenntnistheorie der Modernen Naturwissenschaft* (*The Logic of a Scientific Discovery: On the Epistemology of Modern Science*) (Popper, 1935). Two arguments can be given to support this thesis. First, the idea of falsificationism is, in fact, a very strong blow to the epistemic certainty enjoyed by science, especially the one based on mathematical modeling. Observing the historical development of knowledge, Popper presents a deep epistemological pessimism, noting that there is no possibility of an absolute and certain judgment of the correctness of a theory, condemning any theory to uncertainty in this regard. The only kind of certainty that can be achieved is nonconstructive and only arises when the theory is overthrown. However, this approach is, in fact, a voice for epistemological relativism, for the practical differentiation of the real world and the world of interpretation of reality. The latter world is inherently "arbitrary" in the sense that it is fundamentally deprived of access to the real world and cannot realize the relationship of strict correspondence. It is limited to continuous trials that are always uncertain and ultimately inappropriate.

Second, Popper formulates the conditions of the theory in an extremely interesting and bold way, clearly referring to the assumptions of axiomatic systems in at least two places. First by pointing to the source of a theory which, in his own words, could come from everywhere, from: "an anticipation, a hypothesis, a theoretical system, or what you will" (Popper, 2002, p. 9). The second place is the description of „different lines along which the testing of a theory could be carried out." For this, in the first place the "internal consistency of the system" is examined (*ibidem*), which sounds like a reference to the axiomatic systems of Peano and Hilbert. In his conception, Popper clearly weakens the position of scientific knowledge and even challenges it in a peculiar way, and at the same time, at least partially, frees it from the close relationship with the world. Both these movements seem consistent and open the way to perceiving knowledge, this time scientific, completely different from the image of it that arises with Galileo, who was convinced of discovering the real (mathematical) properties of the world through science.

3. Another premise indicating a change in the understanding of knowledge is provided by Edmund Gettier, the author of the article that shocked the foundations of epistemology (Gettier, 1963). This light tone is justified by the somewhat anecdotal setting of this event and the extremely small volume of the three-page text that accomplished this feat. Using two examples, which actually exhaust the volume of the text, Gettier proves that one can have justified and truthful beliefs, and thus fulfill Plato's conditions

and have no knowledge. In this way, it shows the inaccuracy in reasoning, which leads to the recognition of the hitherto existing definition of knowledge as defective. Admittedly, similar examples are given earlier by Meinong and Russell, but they do not draw so far-reaching conclusions. The latter writes in 1948 that "It is very easy to give examples of true beliefs that are not knowledge," adding that knowledge is a subclass of these judgments: "Every case of knowledge is a case of true belief, but not vice versa" (Russell, 1948, p. 170).

Gettier's text opened a rich and still unfinished discussion on the problem that bears his name. The number of proposals, even exceeding the number of its participants (Borges *et al*., 2018). They propose two main ways of looking for a solution: introduce the fourth condition to the incomplete three proposed by Plato, or focus on the third condition in the original Greek wording, i.e. referring to the concept of logos, and give it stricter, noncontestable character (Moser, 2010). From the point of view of this text, it is worth quoting Sober who summarizes Gettier's problem as follows:

> "The skeptical argument contradicts a fundamental part of our commonsense picture of the way we related to the world around us. Common sense says that people have knowledge about the world they inhabit; the skeptical argument says that common sense is mistaken in this respect." (Sober, 2005, p. 157)

In this statement, he refers to common experience, pointing to a certain banality of the situation described by Gettier, and at the same time, locating it in the world of a certain everyday life completely abandons speculation. In this way, knowledge, which is after all the main subject of reflection, acquires a practical nature.

Following this path and examining the pragmatic foundation of the presented thought, we quickly face the necessity of a deeper examination of the examples of Gettier and others similar to certain stories that must also happen as meta-stories about knowledge. Their meaning results from the special position of their reader (and author), who must have a certain special, higher knowledge, coming from a level higher than the world presented in the examples. This type of perspective cannot be ignored when examining the conditions of the existence of knowledge, which can reveal itself at any level of the analysis and cannot be deprived of participation in the analysis at any of them. Such an observation is, of course, endowing knowledge with a quality such as semantic dispersion and contextual dependence. This reasoning is modeled on Tarski's approach, who distinguished different types of language that was based precisely on the diversification of the levels of its existence. He differentiated the level (and language) of direct talking about the world and the meta-level (metalanguage). The latter makes it possible to establish the rules for the first utterance (Tarski, 1933). There is no place here for a broader analysis of this issue, which I have presented elsewhere,

but it leads directly to the understanding of one of the most important aspects of the decomposition of knowledge, referring directly to the inspiration taken from the crisis of the mathematical description of the world, described earlier and finally to the process of relativizing knowledge.

4. The next, fourth premise of the idea of knowledges concerns the phenomenon both easy and difficult to present. Its apparent ease lies in the specific consistency of assumptions referring to the social context of the described phenomena, the description of which proves to be difficult due to their quantity, variety, and interrelationships. The latter difficulty is also paradoxically helpful; reflection in this area is extensive and well-known, so in this text it is enough to recall the most important facts. The perspective of knowledge as a hostage of social circumstances, as Marian Adolf and Nico Stehr claim, appears in the writings of classics researching social reality, such as Max Weber, Max Scheler, Karl Marx, Karl Mannheim, Georg Simmel, and Emil Durkheim (Adolf, Stehr, 2014). Among them, it was Scheler who in 1924 proposed the idea of the sociology of knowledge, proposing the appropriate concept: *Wissenssoziologie*. Adolf and Stehr, however, nominate Karl Mannheim as the proper father of the sociology of knowledge, who devotes a separate chapter to it in his work from 1929 entitled *Ideology und Utopie* (Ideology and Utopia). The emblematic representative of this direction and its further development is David Bloor, one of the co-founders and main representatives of the so-called *strong programme*, i.e. a research school, named after the place of its inception the Edinburgh School. He presents his views in a book *Knowledge and Social Imagery*, published in 1976, which is still a kind of key reference for the sociology of knowledge. It includes the following sentence: "instead of defining it [knowledge] as true belief—or perhaps, justified true belief—knowledge for the sociologist is whatever people take to be knowledge. It consists of those beliefs which people confidently hold to and live by" (Bloor, 1991, p. 5), openly contesting the Platonic conditions of knowledge to which he directly refers.

The sociology of knowledge, as part of the project of its social conditioning, is supplemented by the sociology of science, which can also be understood as a field of realization of scientific knowledge, i.e. knowledge that is a subject of the methodological conditions of its correctness. These conditions, stable until the end of the nineteenth century, based primarily on the certainty of mathematical judgments, turned out to be questioned. A rich trend of reflection in this field is opened by Ludwik Fleck, whose concepts, presented in 1936, are an inspiration both for Thomas Kuhn and his ideas of scientific revolutions and for Bruno Latour. Latour also, as Sady points out, nominates Fleck as the father of the sociology of science (Sady, 2013, p. 211). Kuhn represents a very general type of approach, trying to synthesize the processes of the development of science, approaching the philosophy of science. Paul Feyerabend and Imre Lakatos share a similar, constructive ap-

proach. The views of all three constitute mainly a phenomenon described as a "historical turn" in the understanding of science (Bird, 2008).

The trend represented by Latour is much closer to the social reality and the institution belonging to it, which is the laboratory, i.e. a place specially constructed for acquiring knowledge. This is the direction of publications of Robert Merton from 1973 (Merton, 1973), Steven Woolgar and Bruno Latour from 1979 (Latour, Woolgar, 1979), or Karina Knorr-Cetina from 1981 (Knorr-Cetina, 1981). This research develops, creating its own field of reflection called Science and Technology Studies (STS), the basic assumptions of which were formulated in 1991 by Steve Woolgar (Woolgar, 1991). Latour, who, as he writes about himself, developed the sociology of science in the 1980s, ultimately built an extremely important and extensive social theory, going far beyond the strict field of science or knowledge. This type of approach, representing a philosophical approach to the social and historical circumstances of the existence of knowledge, is also represented by philosophers such as Foucault or Lyotard. This multi-threaded, extensive and different reflection on knowledge, also in its scientific embodiment, is at the same time an extensive story about the whole of society and therefore also its political, economic, and anthropological contexts. It is also a great break with the idea of knowledge, empowered transcendentally or metaphysically. Instead, knowledge is reduced to its numerous, dispersed, often not obvious and surprising, but permanently present, social contexts, leading to its diversification.

5. The contexts mentioned in the previous paragraph lead us immediately to the next, fifth premise of the reasoning presented in this text. It consists of three elements, or more precisely, three different articulations of the concept of knowledge that appears as a term in three different areas of reflection and use. These reflections, however, have a general common feature and differ only in the scale or nature of their implementation. All of them treat knowledge in a reified, instrumentalized way, perceiving it as a physical entity that can be used and utilized, and thus becomes a part of wider processes. The first reflection is created by the philosophers mentioned here: Foucault and Lyotard, among whom the latter, in particular, sees knowledge as a component of substantive social processes leading to political consequences. He writes about knowledge, which, according to him, should be widely available by endowing it with electronic character but also may constitute the main component of political domination (Lyotard, 1979). This kind of approach continues a slightly earlier concept by Graham Bell in his book *The Coming of Post-industrial Society: A Venture in Social Forecasting* from 1973, which develops the theses presented by Fritz Machlup in 1962. According to what Peter Drucker writes about himself, this approach is initiated by him with a text from 1961. The quoted texts gradually move forward in identifying the social role of knowledge, ultimately making it the

basis of the functioning of society, understanding this conclusion operationally, as a result of its economic and political role, perceived directly as a factor of advantage and power. Knowledge is understood as a resource that gradually plays an increasingly important market role and inevitably becomes a kind of unusual good or even a natural resource.

The political level must therefore be supplemented with a purely technical level of use, at which knowledge becomes the subject of management at the level of an organization, most often an enterprise, which classifies it as one of its more and more valuable resources. This kind of understanding of knowledge, purely instrumental, utilitarian, and practical, appears in the relevant branch of management: knowledge management (KM). Kimiz Dalkir, an author of the textbook *Knowledge Management*, defines it as follows:

> "Knowledge management represents a deliberate and systematic approach to ensure the full utilization of the organization's knowledge base, coupled with the potential of individual skills, competencies, thoughts, innovations, and ideas to create a more efficient and effective organization." (Dalkir, 2005, p. 2)

This definition is somewhat of a compromise, as Dalkir has found over a hundred similar ones, most of which are correct. This situation results from the fact that this management turns out to be extremely heterogeneous and dependent on the research context, which may be surprisingly different in nature. Dalkir gives an impressive list of examples of such contexts:

> "organizational science, cognitive science, linguistics, and computational linguistics, information technologies such as knowledge-based systems, document and information management, electronic performance support systems, and database technologies, information and library science, technical writing and journalism, anthropology and sociology, education and training, storytelling and communication studies, collaborative technologies such as computer supported collaborative work and groupware, as well as intranets, extranets, portals, and other web technologies." (Dalkir, 2005, p. 6)

It turns out, therefore, that the knowledge of an organization can be revealed in different ways what is caused by the way it is understood. On the other hand, pragmatism in the approach to knowledge in this area is not only utilitarian and instrumental, but above all extremely unambiguous: the only goal is effective and orderly use of knowledge. A number of practical methods serve this purpose, the list of which is opened by Nonaki's classic proposal, entitled SECI (Socialisation, Externalisation, Combination, and Internalization) (Nonaka, 1991; Nonaka, Takeuchi, 1991). This is how another knowledge management classic, Karl Wiig, interprets also the role of knowledge. Ultimately, it serves two, not very complicated purposes: it is to make the company function intelligently, making the best use of the knowledge resources it has at its disposal (Wiig, 1997). Wiig formulates

them in a text summarizing the history of knowledge management from 1997. Maximizing efficiency—using the nomenclature proposed by him—is clearly presented there in an appropriate scheme presenting individual main tasks and their operational implementation. Wiig also presents the chronology of the formation of this management, which shows that it developed rapidly in the 1990s and, apart from very early ideas, gradually matured in the 1980s. It is, therefore, undoubtedly a continuation and operationalization of earlier political or philosophical ideas.

The last of the three different contexts in which knowledge appears as a useful object or resource is the area of reflection known as the Knowledge Organization (KO). Its real source is a library in which knowledge is gathered in a tedious cataloging process, allowing access to an extensive, spontaneously accumulating repository. This initially technical process, however, immediately updates questions about the structure of the knowledge and ultimately about its content. The birth of the concept of knowledge organization is reported by Hider who writes: "The term 'knowledge organization' was chosen for the English name to represent wider interests than classification, although these did not at first extend to other IO [Information Organization] activities such as descriptive cataloging: it was abstract rather than recorded knowledge that was to be organized into schemes and vocabularies" (Hider, 2018).

Its mature version is given by the main author of this reflection, Birger Hjørland: "KO is about describing, representing, filing and organizing documents and document representations as well as subjects and concepts both by humans and by computer programs" (Hjørland, 2016). The effect of development is also his specific understanding of knowledge that modifies the subject. Hjørland turns it into an idea he calls "knowledge claims" (Hjørland, 2003, p. 100). This movement allows him to understand, on the one hand, the multitude of interpretations of the world within knowledge, and, on the other hand, the variety of structures organizing this knowledge. The shift of knowledge towards its claims, and then concepts, is, however, an admission of its dispersion and confusion, articulating the most important experience born in the observation of technical stores of knowledge. Ultimately, however, the goal remains the same: it is the exploitation of knowledge, its extraction, and use, which allows the Knowledge Organization to be placed next to the previous approaches, which build the fifth premise of the thesis of relativizing knowledge and its reification at the same time.

6. The last, sixth premise refers to the field of events that are happening intensively today, although they are the result of many years of research, dating back to the mid-twentieth century, and the sources of which are even earlier. It also concerns a very specific field, which is digital technology. These searches are also in the phase of spectacular development, considering the number of publications and solutions appearing for example in the

area of the so-called artificial intelligence. Artificial intelligence is also the traditional and oldest field of knowledge exploitation in information technology (IT), dealing with the problem of the so-called knowledge representation, or more specifically, knowledge representation and reasoning. As Jurfsky and Martin write, the first ideas of this kind appeared as early as 1957 (Jurafsky, Martin, 2020, p. 329), while the famous conference at Dartmouth College in 1956 is considered to be the symbolic birth of artificial intelligence (Flasiński, 2016, p. 4). However, artificial intelligence searches for its philosophical foundations in terms of knowledge much earlier. Russell and Norvig, authors of a classic textbook in this area, write that "Aristotle argued (in *De Motu Animalium*) that actions are justified by a logical connection between goals and knowledge of the action's outcome" (Russell, Norvig, 2010, p. 7), at the same time formulating the most important aspect of understanding knowledge emerging as an issue in the area of artificial intelligence, which is practical and purposeful utility. This opinion is also critically important from the point of view of this paper because it breaks definitively the direct link between knowledge and man, which ultimately leads to appreciating an artificial system, which is a product of technology, a full-fledged disposer of knowledge.

Knowledge representation is a very extensive field. It is based on the assumption that knowledge can be represented using formal structures, for example logical, and this logic can be very various. The proposals that have developed in this field during its many years of development are numerous (Brachman, Levesque, 2004; Van Harmelen *et al.*, 2008). However, as Russell and Norvig write:

> "Much of the early work in *knowledge representation* (the study of how to put knowledge into a form that a computer can reason with) was tied to language and informed by research in linguistics, which was connected in turn to decades of work on the philosophical analysis of language." (Russell and Norvig, 2010, p. 16)

This observation is also valid today, although the approach to language has changed fundamentally.

To understand the importance and type of knowledge used in the most important solutions in the field of artificial intelligence, i.e. deep learning—the technology of artificial neural networks, one can base on the basic characteristics of knowledge presented by Mariusz Flasiński in his textbook (Flasiński, 2016). It breaks down into two basic approaches, the rivalry of which reflects not only the historical development of artificial intelligence but also the hopes and disappointments associated with it. This emotional context is not only anecdotal, but illustrates the special importance of technology that ultimately mimics human action or thinking, or at least rational action or thinking, referring to the characteristics given by Russell and

Norvig (Russell, Norvig, 2010, p. 2). Flasiński points to two historical approaches: the so-called symbolic artificial intelligence and the so-called computational intelligence. In the first one, knowledge is symbolically represented (in the form of graphs, logical formulas, or symbolic rules) and is explicit. In the second approach, the representation of knowledge is numerical and implicit. Knowledge is, as in the connectionist model, distributed in a form of individual numerical values (e.g. weights in artificial neural networks) which cannot be directly interpreted semantically (Flasiński, 2016). Historically, the second approach is older and opens up the history of artificial intelligence, but the symbolic trend has dominated since the 1970s. More or less at the turn of the 20th and 19th centuries, however, it experiences a breakdown and loses its importance due to the restored connectionist approach. This approach is developing rapidly and spectacularly until today, bringing, among others, language models capable of performing complex cognitive operations, such as text understanding, question answering, etc. Such skills are shown by models from the GPT family (Generative Pre-trained Transformer). The latest version: GPT-3 (Brown et al., 2020) currently shows the state-of-art of development in the field of natural language processing (NLP).

Another and equally spectacular field of development of digital technology related to knowledge is the area of data acquisition and analysis, having various technical implementations known under names such as Big Data, data mining, Internet of Things (IoT), etc. These technologies obviously work together in conjunction and create a certain technological universe, combining various solutions that pursue different particular goals of their stakeholders. One of them is information gathering and knowledge gathering. Big data, as Misa Kinoshita and Kijima Mizuno write "represent projections of things on real world, thinking of people, results of calculations of computer" (Kinoshita, Mizuno, 2017, p. 92), it is, therefore, a powerful and ever-growing digital source of knowledge. It is available in the so-called Knowledge Discovery, which is the central task of the technology called data mining. It is, as Bramer writes, "non-trivial extraction of implicit, previously unknown and potentially useful information from data" (Bramer, 2016, p. 2).

A characteristic feature of the modern approach to data is its holistic nature, resulting from the total area of their presence. In other words, data, in the sense given by Kinoshita and Mizuno, cover more and more areas of the world and penetrate deeper and deeper into its processes and phenomena. Of course, for this reason, they also become a source of serious ethical problems (Chandler, Fuchs, 2019). They also raise the question of the type and status of knowledge they become. Insights emerging in this field, for example, knowledge extraction based on large virtual social networks (e.g. Facebook), show its extraordinary diversity, multivariant or even contradiction, activate the need to understand the social processes of its construction and

proliferation, etc. From this point of view, modern technologies of data acquisition and analysis turn out to be great repositories of distributed knowledge, with surprising forms of articulation and not obvious in terms of their relationship with individual people.

## CONCLUSIONS

The historical reconnaissance of the contexts in which the concept of knowledge appears here clearly shows the significant evolution of the idea of knowledge. Today knowledge deviates from its Platonic definition, in particular abandons man as its disposer and abandons the condition of truthfulness. This evolution takes place first as a result of a fundamental change concerning the possibility of representing the world by epistemological human constructions, or more precisely, the rejection of this possibility entirely in favor of arbitrary, free constructions, meeting only the condition of assumed, internal coherence.

The second most important movement leading to the erosion of understanding of knowledge is placing it in the context of social reality, based on the assumed close, mutual relationship. This movement is expressed by a whole range of ideas, that produce the need to reconstruct the conditions of cognition. This need is similar to the one previously described, but this time it is caused in fact by the redefinition of key subjective features. It consists of departing from the essential interpretation in favor of the social one. It is, of course, also a dramatic process and full of numerous consequences, also concerning knowledge, which turns out to be socially determined, which causes its dispersion, decentralization, and eliminates the condition of truthfulness. At the same time, by becoming a hostage of historical and social circumstances, knowledge is reified and interpreted as a resource or good, thus becoming a source of further social and political transformations.

The third and most important variant of knowledge reinterpretation appears as a result of digital technologies. Here, too, knowledge is subject to concretization, which is of a formal or even numerical nature. The approach to it is strictly instrumental, teleological, and utilitarian. Knowledge becomes a local phenomenon and is subject to computational processes.

The accumulated premises justify an evolution in which knowledge is understood as dispersed (distributed). On the one hand, this dispersion concerns the level of reflection, which means that a cognitive event that causes the referent of the concept of knowledge becomes heterogeneous and local because of the interpretative approach adopted. On the other hand, the characteristic of dispersion is also observed at the direct level of articulation of knowledge that does not claim uniformity, which is closely correlated with the removal of the requirement of truthfulness. In this situation, it be-

comes justified to introduce the plural in the modern conception and practice of knowledge, i.e. the concept of "knowledges," which defines the current way of existence of knowledge.

The paper is a result of the realization of the research project number 2018/29/B/HS1/01882 financed from the resources of the National Science Centre, Poland.

## REFERENCES

M. Adolf, N. Stehr, *Knowledge*, Routledge, Abingdon–Oxon 2014.
K. A. Appiah, *Thinking It Through: An Introduction to Contemporary Philosophy*, Oxford University Press, Oxford–New York 2003.
R. Audi, *Epistemology: A Contemporary Introduction to the Theory of Knowledge, 2nd Edition,* 2rd edition., Routledge, New York 2003.
A. Bird, *The Historical Turn in The Philosophy of Science, in: The Routledge Companion to Philosophy of Science,* S. Psillos, M. Curd (Eds.), Routledge, London– etc., 2008, pp. 67–77.
D. Bloor, *Knowledge and Social Imagery*, 2nd Edition, University of Chicago Press, Chicago 1991.
R. Borges, C. de Almeida, P. D. Klein (Eds.), *Explaining Knowledge: New Essays on the Gettier Problem*, 1st edition., Oxford University Press, Oxford, UK 2018.
R. Brachman and H. Levesque, *Knowledge Representation and Reasoning*, 1st edition, Morgan Kaufmann, Amsterdam–Boston, 2004.
M. Bramer, *Principles of Data Mining*, 3rd ed., Springer-Verlag, London 2016.
T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, et al., *Language Models are Few-Shot Learners*, ArXiv:2005.14165 [Cs], 2020.
M. Burgin, *Theory of Knowledge: Structures and Processes*, Vol. 5, World Scientifc Publishing Co. Pte. Ltd., Singapore 2015.
D. Chandler, C. Fuchs, *Introduction Big Data Capitalism—Politics, Activism, and Teory*, in: Digital Objects, Digital Subjects: Interdisciplinary Perspectives on Capitalism, Labour and Politics in the Age of Big Data, Chandler, D. and Fuchs, C. (Eds.), University of Westminster Press, London, 2019.
R. Chisholm, *Theory of Knowledge*, Pearson College Div, Englewood Cliffs, N. J. 1989.
P. F. M. Cornford, *Plato's Theory of Knowledge, The Theaetetus and the Sophist of Plato Translated with a Running Commentary,* Kegan Paul, Trench, Trubner & Co., New York 1935.
K. Dalkir, *Knowledge Management in Theory and Practice*, Elsevier–Butterworth Heinemann, Boston 2005.
R. Descartes, *Meditationes de Prima Philosophia, in: Qua Dei Existentia et Animæ Immortalitas Demonstratur*, Michel Soly, Paris 1641.
M. Deschoux, *Platon Ou Le Jeu Philosophique*, Presses universitaires de Franche-Comté, Le Belles Lettres, Paris 1980.
J. Dewey, *Philosophy & Civilization*, Minton, Balch & Company, New York 1931.
M. Flasiński, *Introduction to Artificial Intelligence*, Springer, Cham, Switzerland 2016.
E. L. Gettier, *Is Justified True Belief Knowledge?*, Analysis, 23 (6), 1963. pp. 121–123.
P. Hider, *The Terminological and Disciplinary Origins of Information and Knowledge Organization*, Education for Information, 34 (2), 2018, pp. 135–161.
D. Hilbert, *Grundlagen der Geometrie*, Verlag von B.G. Teubner, Leipzig, 1899.
B. Hjørland, *Fundamentals of Knowledge Organization*, Knowledge Organization, INDEKS VERLAG, 30 (2), 2003, pp. 87–111.
B. Hjørland, *Knowledge Organization*, Knowledge Organization, 43 (6), 2016, 475–484.
L. H. Hodgkin, *A History of Mathematics: From Mesopotamia to Modernity*, OUP Oxford, Oxford 2005.

D. Jurafsky, J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed., Prentice Hall, 2020.
E. Kinoshita, T. Mizuno, *What Is Big Data*, in: Big Data Management, F. P. García Márquez, Lev, B. (Eds.), Springer International Publishing, Cham, 2017, pp. 91–101.
M. Kline, *Mathematical Thought from Ancient to Modern Times,* Vol. 3, New Ed edition., Oxford University Press, New York, 1990.
K. Knorr-Cetina, *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*, Elsevier Science Limited, 1981.
B. Latour and S. Woolgar, *Laboratory Life: The Construction of Scientific Facts*, Princeton University Press, Princeton, N.J, 1979.
J.-F. Lyotard, *La condition postmoderne. Rapport sur le savoir*, Éd. de minuit, Paris 1979.
J.-F. Lyotard, *Le Différend*, Éd. de minuit, Paris 1983.
R. Maciąg, *Transformacja Cyfrowa. Opowieść o wiedzy*, TAiWPN Universitas, Kraków 2020.
R.K. Merton, *The Sociology of Science: Theoretical and Empirical Investigations*, University of Chicago Press 1973.
P. K. Moser, *Gettier Problem*, in: A Companion to Epistemology, J. Dancy, E. Sosa, E., M. Steup (Eds.), Wiley-Blackwell, Malden, 2010, pp. 395–397.
R. Murawski, *Filozofia Matematyki. Zarys Dziejów*, 2nd ed., Wydawnictwo Naukowe PWN, Warszawa 2001.
R. Murawski, *Essays in the Philosophy and History of Logic and Mathematics*, Editions Rodopi B.V., Amsterdam 2010.
I. Nonaka, *Knowledge-creating Company,* Harvard Business Review, 69 (6), 1991.
I. Nonaka and H. Takeuchi, *The Knowledge-Creating Company How Japanese Companies Create the Dynamics of Innovation*, Oxford University Press, New York, 1991.
G. Peano, *Arithmetices principia: nova methodo*, Fratres Bocca, Romae–Florentiae 1889.
C. S. Peirce, *What Pragmatism Is*, *The Monist*, 15 (2), 1905, pp. 161–181.
S. R. Penrose, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics,* Oxford University Press, Oxford, 1999.
Plato, *Platons Werke*, translated by Schleiermacher, F., Georg Reimer, Berlin, 1856.
Plato, *Theatetus*, R. H. Waterfield (trans.), Penguin Classics, Harmondsworth, Middlesex, England– New York 1987.
Platon, *Parmenides,* Teajtet, W. Witwicki (trans.), Antyk, Kęty 2002.
K. Popper, *The Logic of Scientific Discovery: On the Epistemology of Modern Science*, 2nd edition, Routledge, London 2002.
D. Pritchard, *What Is This Thing Called Knowledge?*, Routledge, London 2006.
B. Russell, *Human Knowledge—Its Scope And Limits*, George Allen and Unwin Ltd., London 1948.
S. J. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach*, Pearson, New Jersey 2010.
W. Sady, *Spór o Racjonalność Naukową. Od Poincarego Do Laudana*, Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, Toruń 2013.
E. Sober, *Core Questions in Philosophy. A Text with Readings*, 2005.
A. Tarski, "Pojęcie prawdy w językach nauk dedukcyjnych", *Prace Towarzystwa Naukowego Warszawskiego*. Wydział III Nauk Matematyczno-Fizycznych, VII (34), 1933.
F. Van Harmelen, V. Lifschitz, B. Porter (Eds.), *Handbook of Knowledge Representation*, Elsevier, Amsterdam–Boston 2008.
K. M. Wiig, *Knowledge Management: Where Did It Come from and Where Will It Go?, Expert Systems with Applications*, 13 (1), 1997, pp. 1–14.
J. Woleński, *The History of Epistemology*, in: Handbook of Epistemology, I. Niiniluoto, M. Sintonen, J. Woleński (Eds.), Springer Netherlands, Dordrecht 2004, pp. 3–54.
S. Woolgar, *The Turn to Technology in Social Studies of Science*, Science, Technology, & Human Values, 16 (1), 1991, pp. 20–50.

ABOUT THE AUTHOR — PhD, associate professor, Jagiellonian University, Institute of Information Studies.

Email: rafal.maciag@uj.edu.pl

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Pavel N. Baryshnikov

# EXTENSION OF CRITICAL PROGRAMS OF THE COMPUTATIONAL THEORY OF MIND

doi: 10.37240/FiN.2022.10.zs.12

***ABSTRACT***

Technological advances in computer science have secured the computer metaphor status of a heuristic methodological tool used to answer the question about the nature of mind. Nevertheless, some philosophers strongly support opposite opinions. Anti-computationalism in the philosophy of mind is a methodological program that uses extremely heterogeneous grounds for argumentation, deserving analysis and discussion. This article provides an overview and interpretation of the traditional criticism of the computational theory of mind (computationalism); its basic theses have been formed in Western philosophy in the last quarter of the 20th century. The main goal is to reveal the content of the arguments of typical anti-computationalist programs and expand their application to the framework of the semantic problems of the Classic Computational Theory of Mind. The main fault of the symbolic approach in the classical computationalism is the absence of a full-fledged theory of semantic properties. The relevance of considering these seemingly outdated problems is justified by the fact that the problem of meaning (and general problems of semantics) remains in the core of the latest developments in various areas of AI and the principles of human-computer interaction.

**Keywords**: anticomputationalism, computational theory of mind, Chinese room, finite automata, symbolic semantics, language of thought.

## CRITICAL PROGRAMS OF COMPUTATIONALISM: A CLASSIC SET OF ARGUMENTS

This article analyses critical programs of various forms of the computeristic paradigm at various stages of its formation. The main goal is to identify key methodological trends in critical programs, to supplement existing classifications and consider possible responses from modern cognitive sciences and the engineering theory of artificial intelligence.

Today, review papers on various versions of the computational theory of mind offer the so-called typical list of critical programs:

— Variations of John Searle's Chinese Room argument,
— the triviality argument,
— Kurt Gödel's incompleteness theorem argument,
— the limits of the computational modeling argument,
— the temporal argument,
— the embodied cognition argument.[1]

Let us consider in more detail each of the above-listed positions, accompanied by our own commentary and interpretation.

## The Chinese Room Argument

Essentially, this argument is based on three assumptions: 1) computations have the properties of multiple realization, whereas the mental does not; 2) intentionality, as opposed to a computer program, is determined by means of content, not syntactic structures; 3) a program is not a product of a computer (it is written by an intelligent coder), while the mind is generated by the brain, where the content of the mind is presented. It is characteristic that Searle's arguments correlate with the problem of information ontology in computing systems, where the translation of regular syntax into arbitrary semantics and vice versa remains extremely relevant.[2]

Today, computationalists have counterarguments based on the newest advances in computer technology only for the judgment number 3. It should be noted that the cyclical and recursive structures of algorithms, in which a program generates other programs, have been known for a long time. Most often, their use was associated with the translation of expressions from a programming language into a low-level language of machine instructions. Today, there is a type of automatic machine learning, when algorithms themselves write sub-programs and train neural networks. We should mention that the code created by a machine surpasses the code written by a programmer in a number of parameters.[3] Of course, the artificial program generation has nothing to do with the products of the mental realm. However, here the very structure of the so-called machine supervisor (observer) is important, during its functioning the autonomy of the entire machine system's behavior increases. As the result of this increasing the complexity of the processed data and the relevance of responses to the user or the physical environment requests increase. All this brings machine intelligence closer to the functional benchmark of the Turing test. In this case, the functional sig-

---

[1] M. Rescorla, *The Computational Theory of Mind*, in: The Stanford Encyclopedia of Philosophy, E. N. Zalta (Ed.), 2017; https://plato.stanford.edu/entries/computational-mind/

[2] J. R. Searle, *Minds, Brains, and Programs*, Behavioral and Brain Sciences, 3 (3), 1989, pp. 417–424.

[3] T. Simonite, *AI Software Learns to Make AI Software*, 2017; https://www.technologyreview.com/2017/01/18/154516/ai-software-learns-to-make-ai-software/

nificance of the fact that the machine does not think is leveled—an adequate action is sufficient. There are statistical models of natural language that calculate the probabilities of the distribution over the sequence of tokens. After the stunning results in the generation of natural language texts obtained during the work on the GPT-3 autoregressive transformer, the arguments based on the principle of the "Chinese room" argument can be considered untenable. GPT-3 generates coherent human-like texts by extracting content from a vector representation of a gigantic number of sequences. "GPT-3 was writing articles, penning poetry, answering questions, chatting with lifelike responses, translating text from one language to another, summarizing complex documents, and even writing code."[4] The issue of syntax translating into semantics becomes irrelevant. The issue of the connection between meaning formation and probabilistic statistical models becomes more significant.

## Triviality Argument

Triviality as an extreme degree of simplification in describing a certain system underlies the thesis, which is often found in the works of anti-computationalists of the mid-20th century. Essentially, this thesis is about the fact that any physical process can be represented as a computational function, since a quantitative measure is applicable to all properties of matter. "Every ordinary open physical system implements every finite-state automaton."[5] An open system is understood here as a system that has continuous interaction with the environment. This interaction can take the form of information, energy or material transformations due to the permeability of the system boundary. Thus, the Classical Computational Theory of Mind is recognized as trivial, since it represents mind as an open system described by the functions of a finite-state machine. In a simplified form, an abstract finite-state machine starts its operation from the initial state and then changes its internal states in accordance with the transition function. The transition function is defined in terms of the set of states that can be transitioned from the current state. It is important that this set is finite. The admissibility of a transition is determined by regular events that correspond to a finite set of internal states. It is obvious that the complexity level of the interaction between the organism and the environment (not to mention the meaningful properties of mental states) exceeds the executive capabilities of the finite automaton. Computationalists' counter-arguments are based on various modifications of the computational theory, including semantic,

[4] S. Tingiris, B. Kinsella, *Exploring GPT-3*, Packt Publishing Ltd., Birmingham 2021, p. 3.

[5] H. Putnam, *Minds and Machines*, in: Dimensions of Mind, S. Hook (Ed.), New York University Press, New York 1960, pp. 148–180.

causal, and other aspects. In other words, the triviality argument sounds like this: if any physical systems are described by the properties of a finite automaton, then computationalism is trivial; if not, then computationalism is not complete.

## Gödel's Incompleteness Theorem Argument

Considering that a voluminous set of works is devoted to this problem, let us give a brief (interactive) description of the argument and counter answers.

— Anti-computationalists: human mathematical abilities are superior to the computational capabilities of a Turing machine, because a person is able to understand the meaning of the Kurt Gödel's incompleteness theorem[6] (Lucas, 1961; Nagel et al., 2001).

— Computationalists (and philosophers supporting this position): the anti-computationalist understanding of the Gödel's formal systems incompleteness theorem is based on mathematical errors and false premises.

For example, in his comments to Roger Penrose's *Shadows of the Mind*, David Chalmers points to the absence of a direct connection between the Gödel's argument and the non-computability of physical elements in the theory of quantum gravity.[7] Chalmers emphasizes that if each physical component of the brain has a finite number of relevant states, then these causal relations between the states of the brain are representable in a discrete computational form, despite the continuity of natural processes. In other works, he denies the persuasiveness of external and internal critical arguments and points to the universalism of computational models in the reproduction of the causal structure of the mental.

## Limits of Computational Modeling Argument

This anti-computeristic "line of defense" is built on the intuitive assumption that there are many aspects in human activity that go beyond the explanatory capabilities of formal systems: creativity, development, understanding, heuristics, planning, etc. The rigid logical limitations of computer models do not allow reflecting the flexibility, stability, and adaptability of many cognitive processes. Criticizing the Classical Computational Theory of Mind, Jerry Fodor points out that Turing-type computer modeling explains

---

[6] J. Lucas, *Minds, Machines and Gödel*, Philosophy, 36, 1961, pp. 112–127; E. Nagel, J. R. Newman, D. R. Hofstadter (Rev., eds.), *Gödel's Proof*, New York University Press, New York. 2001, p. 129.

[7] D. J. Chalmers, *Facing up to the Problem of Consciousness*, Journal of Consciousness Studies, 2, 1995, pp. 200–219.

only local fragmentary processes and is not able to adequately represent the abductive elements of probabilistic knowledge and conditional experience.[8] Note that when criticizing the early version of the language of thought hypothesis, Fodor's follower S. Schneider opposes to his arguments the idea that, firstly, Turing computations are sensitive to the properties of an integral system, and secondly, abductively derived knowledge is computable within the framework of a formal pragmatists.[9]

## Temporal Argument

The key thesis defended by this argument is that mental processes take place over time. In addition, the human mind is capable of solving complex problems in a non-trivial way. The main emphasis of the proponents of this argument is that the classical model of sequential computation cannot cope with the explanation of the temporal characteristics of cognition processes. An abstract Turing machine does not take into account the resource constraints (time and energy) imposed on computations by the physical world. It is important to emphasize that this argument is used not only by anti-computationalists but also by proponents of alternative computational approaches.[10]

Various types of neural models or parallel computing are proposed to reproduce the elements of mental processes. It is argued that an abstract computational model can be equipped with temporal properties because each discrete step of computation can formally correspond to a certain moment in time or other physical parameter. The technological interpretation of the representation is of interest. A computer is an artificial system capable of correlating scalar values with analog physical signals. For example, when digitizing sound, the compressed waves are sent to a transducer (microphone diaphragm), which transmits them in the form of voltage fluctuations. These fluctuations are then encoded into a digital bit rate (the number of bits used to transmit data per unit of time). It turns out that at any moment of time the state of the system represents the spatio-temporal states of physical waves.

However, technological comparisons of the operation of neurons with an analog-to-digital converter (ADC) inherit the entire set of engineering complexities. The imperfection of the ADC is due to the fact that when digitizing the analog signal's continual function of the time, distortions and errors are inevitable, to which the limitation of the frequency spectrum is added.

---

[8] J. A. Fodor, *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology,* MIT Press, Cambridge, Mass. 2000, p. 126.

[9] S. Schneider, *The Language of Thought: A New Philosophical Direction,* MIT Press, Cambridge, Mass.–London., 2011.

[10] G. Piccinini, *Physical Computation: A Mechanistic Account*, Oxford University Press, Oxford 2015.

Therefore, according to the Kotelnikov-Shannon theorem, a complete digital restoration of an analog signal is impossible.[11] Thus, any digital representation of analog physical processes is always an approximation.

There are analytic computational counterarguments of the following type: the fact that the physics of cognitive processes has continuity does not mean that computational models must include this continuity. Physical states exist in continuous time, but this is not reflected in any way in the digital logic of the device itself. As a result, the idea is substantiated that computational models of mind should not describe absolutely all physical processes in the brain. The question remains open as to whether the continuity of physical processes entails the necessary continuity of cognitive processes. This is not obvious to the supporters of computationalism.

## Embodied Cognition Argument

The concept of embodied cognition was formed as a result of the perceptual studies of Maurice Merleau-Ponty and James J. Gibson, who developed an ecological approach. The essence of this approach is to study the unity of cognition and bodily action "built-in" into the challenges of the environment. With this formulation of the issue, the processes of mind cannot be regarded as abstract manipulations of symbols. Computationalism is opposed by environmentalism, in which the unity of mind, body, and environment is described in terms of the theory of dynamical systems.[12] Computationalists, in turn, argue that the computational approach is sufficient to represent the dynamic relations of the organism and the environment in the form of a system of incoming signals and outgoing motor-communicative actions.

Summarizing the so-called typical set of critical anti-computational programs, we can point out that, despite the variety of arguments, they are all united by the same logic of reasoning. This logic is based on the following principle:

— mental processes are derived from physical ones,
— physical processes causative of mental content have computational properties,
— physical processes in computational models do not cause mental content,
— therefore, computationalism is false, or at least incomplete (does not explain all the variety and complexity and mental content).

In our opinion, the main problem here is that the methodology of computationalism as such is criticized, not its particular applications. The heuristici-

[11] V. A. Kotel'nikov, *On the Throughput of "Ether" and Wire in Telecommunications,* Uspehi fizicheskih nauk, 176 (7), 2006, p. 762 (in Russian).

[12] F. J. Varela, E. Thompson, E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience*, MIT Press, Cambridge, Mass. 1991.

ty of the computer metaphor in cognitive sciences and the philosophy of mind is so high that it is rather difficult to identify the weaknesses of computationalism in general. Criticism can be strengthened by narrowing the methodological field.

## ANALYTICAL CRITICISM OF THE COMPUTATIONAL THEORY OF MIND

Next, we present the arguments aimed precisely at the classical Fodor's program of the Computational Theory of Mind in the framework of the provisions of the language of thought hypothesis (LOTH). The term "analytical" in the title of the paragraph is associated with the method of conceptual analysis of the main criticism arguments, which is used by a number of authors.[13] Some examples from the "typical list" also overlap with analytical criticism. The key problem here is related to the semantic properties of mental states. Therefore, special attention is paid to clarifying the meaning of the terms used and the contexts of their use. The specificity of the analytical approach also lies in the fact that the thematic area shifts from the problem of computability and logical representability to the problem of meaning, which inevitably brings the research focus to the field of philosophy of language.

Here, it is important to point out the key concept that will be used in constructing the criticism of the CTM—derived intentionality. Intentionality in modern analytical philosophy is interpreted very broadly. In this context, intentionality is understood as the inherent ability of mental states to be aimed at some object or some content. Derived intentionality is understood as the content of linguistic expressions inherited from the primary intentional states of mind used for purposes that lie outside the propositional content of the expression.

There are two different lines of the CTM criticism, but both use the concept of derived intentionality. Each of these critical lines presents difficulties for the computational approach, but the essentially these difficulties differ. The first line is Causal Derivation Objection. The problem with the CTM is that the intentionality of linguistic symbols (prescriptions, illocutionary acts) causally depends on the intentionality of mental states and acts of meaning assignment. The second line of criticism is Conceptual Dependence Objection, which puts forward the following thesis: the conventional concept of "symbolic meaning" conceptually depends on the concept of "mental meaning," which has an internal a priori content.[14] Thus, the first line states

---

[13] S. W. Horst, *Symbols and Computation. A Critique of the Computational Theory of Mind*, Minds and Machines, 9 (3), 1999, pp. 347–381; K. M. Sayre., *Intentionality and Information Processing: An Alternative Model for Cognitive Science*, Behavioral and Brain Sciences, 9 (1), 1986, pp. 121–138.

[14] S. W. Horst, *Symbols and Computation* ..., op. cit., p. 354.

that there is nothing inherent in the semantic properties of symbols, depending on mental states, representations, and discursive symbols. The second line, on the contrary, indicates two types of meanings (symbolic and mental), which cannot be reduced to the same ontology. Let us take a closer look at each of the criticisms.

## Causal Derivation Objection

Causal intentionality is a problem that constitutes the most popular criticism of the CTM. A typical representative of this type is the Chinese Room argument. The semantic properties of intentionality can be inherent in both mental states and linguistic tokens (inscriptions, illocutionary acts). The illocutionary act expresses the semantic characteristics of the mental state. This expression takes place when the speaker performs the act of assigning a meaning that fills the sounds of speech or forms of writing with the content of the intentional state. Intentional causation is possible when the speaker's intention (aimed at making the tokens express a state) causes the utterance to have intentionality. In fact, in order to realize a causal explanation of language tokens, it is necessary to distinguish between two states of the speaker:

1. mental state expressed by a linguistic act;
2. an intentional act, by which the content of this mental state is communicated to the spoken sounds.

These clauses contain inconsistencies for the CTM. The semantic properties of symbols are causally derived from mental states, although the semantic properties of mental states are not derivatives. Therefore, it would be false to explain the semantics of mental states through the semantics of symbols, because:

1. The semantics of mental states is not derivative.
2. Any explanation of significant (relevant) symbols requires an explanation of the semantic properties of symbols, which in turn require an explanation of the semantic properties of mental states.

The argument consists of two statements and looks very convincing:[15]

A. All symbols with semantic properties have to have these properties derivatively.
B. None of the semantic properties of mental states is derivative.

The latter statement directly contradicts the CTM, which states that the semantic properties of mental states are derived from the semantic properties of mental representations.

---

[15] S. W. Horst, *Symbols, Computation, and Intentionality: A Critique of the Computational Theory of Mind*, CreateSpace, Charleston, SC 2011.

Taking an extreme position, Daniel Dennett argues that the semantic properties of high-level cognitive processes are derived from low-level cognitive states (from the intentions of genes).[16] Objection A looks surmountable if we prove that all tokens (inscriptions, sayings, computer symbols) are derivative. Fodor points out that the only way symbols can acquire semantic properties is by inheriting dependence on certain entities that have meaning, or as a result of the act of assignment. However, if Searle asserts that the semantic properties of speech causally depend on the intentional states of the speaker, then Fodor points out that the semantic properties of symbols of mentalism (the language of thought) are inherent. Consequently, the symbols of mentalism have a special nature, different from the symbols on the tape of the Turing machine. This raises the question of the nature of mental computation and its relationship to traditional computationalism.

## Conceptual Dependence Objection

This objection is based on the violation of the identity of terms in the analysis of the semantic properties of mental states and symbols. The terms "intentionality," "semantics," "meaning," "reference" are used concurrently both when discussing the semantic properties of mental states and when discussing the semantic properties of symbols. However, these terms may have different content, which depends on the context of the subject area. In the expression "A means ..." the verb "mean" will have certain content if A is a mental state and different content if A is a symbol. Horst connects such a vague semantics with the paronymy of terms, giving examples of the below type:

(1) *Healthy body / Healthy food.*

or

(2) *Many of John's thoughts have been about Mary of late. / The inscription of the name "Mary" in John's dairy are about Mary.* [17]

It is interesting to consider an example which demonstrates the differences in the semantic intentional properties of mental states and symbols through the contextual differences in the meaning of the preposition "about."

The term "derived intentionality" has similar paronymy. Derived intentionality for symbols (especially in the computer memory tape) does not overlap with the intentionality of mental processes. GO symbols can be interpreted within the lexical convention of the English language (in which case it would be the verb of movement "to go"); can be interpreted in the convention

---

[16] D. C. Dennett, *Consciousness Explained*, Penguin, London 1991, p. 74.

[17] S. W. Horst, op. cit., *Symbols and Computation ...*, 1999, p. 350.

of the Japanese language (then it will be a noun meaning a board game). Although, it is worth pointing out that here we only talk about phonetic symbols, since graphically in the Japanese convention, the game should be indicated by the hieroglyph 碁. Without interpretive conventions (compilation algorithms), these symbols—GO or 碁—mean nothing.

Thus, the CTM semantic problem has two separate interpretations: in terms of mental states and in terms of symbolic operations. In this case, it is important for the CTM proponents to indicate in which interpretative convention the term "meaning" is used and whether it relates to the content of mental states. As a result, conceptual dependency can be expressed by the following simple statement:

*Concept X is conceptually dependent on concept Y only if an adequate analysis of X includes the mention of Y.*

Based on the conceptual dependence thesis, S. Horst summarizes the provisions of his criticism of the CTM approach to intentionality in ten statements:[18]

1. Semantic terms like "intentionality," "semantics," "meaning," "reference" are paronymic and used in different meanings in relation to mental states and symbols.
2. It is necessary to distinguish between the ways of using these terms in relation to the semantic properties of mental states and symbols.
3. Expressions applicable to the semantic properties of symbols are conceptually dependent on expressions applicable to the semantic properties of mental states.
4. Analysis of the attributes of the semantic properties of symbols reveals this dependence, since in the CTM, the semantic properties of symbols refer to the semantic properties of mental states.
5. Any attempt to represent the semantic properties of mental states in terms of the semantic properties of symbols will regress and form a vicious circle.
6. When the CTM claims that mental representations have semantic and syntactic properties, the question arises whether it is about (A) the semantic properties of mental states, (B) the semantic properties of symbols, or special computational semantics (C).
7. Acceptance of interpretation A does not make sense.
8. Acceptance of interpretation B leads to regression.
9. Accepting C, we get convinced that there is no adequate theory of semantic properties in the CTM.
10. The explanatory weakness of the CTM stems from an unclear "vocabulary" of semantic terms.

[18] Ibidem, pp. 354–355.

It is necessary to point out that Horst's argumentation is further strengthened by the fact that a symbol in the traditional semiotic sense differs from the abstract symbols that are manipulated by the "printer" on the Turing machine tape. Here we go back to the metaphorical origins of computationalism. A semiotic symbol in any sign system is a conventional designation of a concept, idea or phenomenon, the content of which is attributed conventionally. In computing systems, a symbol is comparable to a quantitative representation of information. That is, symbolic elements (for example, in the ASCII standard) are structural units of information, the semantic properties of which are exhausted by its specification reflecting functions, information and control relations. At the same time, a symbol in the computing system does not indicate anything other than its own functional properties. What is the specificity of the computer program symbols semantics, and to what extent is this semantics comparable to the semantics of mental states? In other words, does the CTM have sufficient explanatory power in mind-related issues?

In computer science, semantics is the meaning of an abstract syntax (sic!), expressed in terms of a rigorous mathematical model. Semantics in one case represents the set of admissible transformations over the syntactic model. For example, the compiler translates the program language into an equivalent machine language description. In another case, semantics is a description in the metalanguage of permissible transformations, as, for example, in the case of the line-by-line work of the interpreter.

The essence of the computer programs semantics is to create rules for assigning values to the symbolic components of these programs. The specificity of the semantic properties of computer symbols is expressed in the definition of some effectively computable relation as a denotation. This is a basic prerequisite for the adequate functioning of a computer. There are three types of programming languages semantics, the properties of which are really difficult to compare with the content side of mental states (operational semantics, propositional semantics, and denotational semantics). The semantic properties of computer symbols are reduced to the consistent computation of syntactic structures within the constraints of computation theory. Due to the fact that computations are carried out on the physical components of machines, hardware restrictions related to the amount of RAM, processor clock frequency, physical time, etc. are imposed here too.

Thus, we come to the conclusion that the main function of symbolic semantics is the consistency of program syntax with opcodes, addressing modes, and numeric equivalents that will be implemented in the physical states of the machine. Despite the fact that this definition intersects with the functionalist interpretation of mind/brain, it is necessary to recognize the following: if we recognize the content side of phenomenal experience, then the computational understanding of semantics in this matter looks useless.

Machine procedures do not have phenomenal content; therefore, the truth of semantic computations depends not on conformity to the "extra-linguistic" world but on the consistency of syntax within the framework of computation theory. If cognitive sciences have not yet discovered a strong relationship between the limitations of formal models and the limitations of cognitive processes of the brain and mind, then computationalism in a broad sense may be considered a heuristic but not a universal scientific metaphor.

## REFERENCES


D. J. Chalmers, *Facing up to the Problem of Consciousness*, Journal of Consciousness Studies, 2, 1995, pp. 200–219.

D. C. Dennett, *Consciousness Explained*. Penguin, London 1991.

J. A. Fodor, *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*, MIT Press, Cambridge, Mass. 2000.

S. W. Horst, *Symbols and Computation a Critique of the Computational Theory of Mind*, Minds and Machines, 9 (3), 1999, pp. 347–381.

____ , *Symbols, Computation, and Intentionality: A Critique of the Computational Theory of Mind*, CreateSpace, Charleston, SC 2011.

V. A. Kotel'nikov, *On the Throughput of "Ether" and Wire in Telecommunications,* Uspehi fizicheskih nauk, 176 (7), 2006 (in Russian).

J. Lucas, *Minds, Machines and Gödel*, Philosophy, 36, 1961, pp. 112–127.

E. Nagel, J. R. Newman, D. R. Hofstadter (rev., eds.), *Gödel's Proof*, New York University Press, New York 2001.

G. Piccinini, *Physical Computation: A Mechanistic Account*, Oxford University Press, Oxford 2015.

H. Putnam, *Minds and Machines*, in: *Dimensions of Mind*, Hook, S. (ed.), New York University Press, New York 1960, pp. 148–180.

M. Rescorla, *The Computational Theory of Mind*, in: *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (Ed.), 2017, Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/computational-mind/

K. M. Sayre., *Intentionality and Information Processing: An Alternative Model for Cognitive Science*, Behavioral and Brain Sciences, 9 (1), 1986, pp. 121–138.

S. Schneider, *The Language of Thought: A New Philosophical Direction*, MIT, Cambridge, Mass.–London 2011.

J. R. Searle, *Minds, Brains, and Programs*, Behavioral and Brain Sciences, 3 (3) 1980, pp. 417–424.

T. Simonite, *AI Software Learns to Make AI Software*, 2017; https://www.technologyreview.com/s/603381/ai-software-learns-to-make-ai-software/

S. Tingiris, B. Kinsella, *Exploring GPT-3*, Packt Publishing Ltd, Birmingham 2021.

F. J. Varela, E. Thompson, E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience*, MIT Press, Cambridge, Mass. 1991.



ABOUT THE AUTHOR — Doctor of science, Philosophy (the degree equivalent to habilitation), professor of the Department of Historical Socio-Philosophical Disciplines, Oriental Studies and Theology, Pyatigorsk State University, 357532, Russian Federation, Pyatigorsk, Kalinin Avenue 9, Russia.

Email: pnbaryshnikov@pglu.ru

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Robin K. Hill

# A CAUTION AGAINST THE ARTIFICIALISTIC FALLACY

doi: 10.37240/FiN.2022.10.zs.13

***ABSTRACT***

The casual justification of the *influence* of a technology, particularly artificial intelligence, by appeal to the *existence* of the technology constitutes an artificialistic fallacy, analogous to the naturalistic fallacy that is well-known in philosophy. Similar to an invocation of nature to provide moral warrant (the naturalistic fallacy), modern tech evangelists invoke the burgeoning of hardware and software products in order to promote that burgeoning (the artificialistic fallacy). This fallacy is often tacit or committed by omission. Emerging ethical initiatives emphasize the refinement, explanation, and oversight of AI products rather than their fundamental ethical effect, making the fallacy recursive.

**Keywords**: philosophy of computing, ethics of computing, artificial intelligence.

## PHILOSOPHY FOR TECH

Conscientious philosophical analysis can reveal latent presumptions that invite actual and potential problems, in technology as well as other realms. Although the humanities play a minor role at most technology firms (where "minor" may be generous), it is important for professionals in computing technology to understand the context and stance of their work as it relates to human life, now and in the future. We need to detect misconceptions, misleading perspectives, and suppressed issues. While all computing professionals hypothetically have the best intentions, and most of us surely do, it takes extra thinking to perform the ethical analyses that are called for by technological advances. Indeed, that's the point here—that technological advances call for ethical analyses, and in a particular way that is not clearly addressed by myriad contemporary initiatives to bring more ethics to bear on technology. High tech is so successful, by many measures, that it seems to create a mandate to continue on its current trajectory. If that mandate is based on the state of the world, the needs of the consumer, and the utility

provided, that's fine. But if it is based on a subtle transition from the descriptive to the prescriptive, we should object.

## THE NATURALISTIC FALLACY

Philosophers exposed the subtlety in question long before high tech came along. We look briefly at two accounts, by David Hume and by George E. Moore. Hume noticed, in his reading of academic philosophers, that they tended to slide from "is" to "ought;" that is, they tended to use a description of how things are in nature to justify how things should be in human affairs (Hume, 1739, Book 3 Part 1, Section 1). He pointed out that there is no derivation in logic that justifies reaching a normative conclusion from empirical evidence. A couple of centuries later, Moore wrote a well-known exposition of this fallacy, stating it particularly in terms of the theory of evolution: "This is the view that we ought to move in the direction of evolution simply because it is the direction of evolution" (Moore, 1993). Under both of these analyses, people are free to derive their morality from nature, or from some other facts about the world, but they must adopt that as a premise in order to avoid the fallacy and achieve validity in the reasoning.

*Naturalistic Fallacy*: *the derivation of "ought" from "is," with "is" taken as Nature.*

Overt Expression: **"**Vegetarianism can't be morally justified—we've been eating meat for millenia."

Covert Expression: **"**You can't get in the way of progress."

Moore uses the device of the Open Question to expose instances of the naturalistic fallacy: Given that X is the case that holds (in nature), can we still ask whether X is good? If the question whether X is good is a coherent one, then the goodness of X does not follow from its veracity; that is, its status as truth in the world does not make it a moral imperative (ibidem).

To be precise about our interpretation of the Naturalistic Fallacy:

1. The assertion "A is natural, and we ought to A" is coherent, and may be true.
2. The argument "A is natural and we ought to do what's natural, and so we ought to A" is valid.
3. The assertion "A is great, and we ought to do A" is coherent, and may be true.
4. But the argument "A is natural, and so we ought to A" is fallacious.

We object only to the last, the implicit appeal to "how things are" to *justify* how things are (as well as the appeal to "how things are not" to *justify* how things are not).

## THE ARTIFICIALISTIC FALLACY

We identify a similar error, the Artificialistic Fallacy (so named because instances appear particularly in comment on artificial intelligence technology).

*Artificialistic Fallacy*: *the derivation of "ought" from "is", with „is" taken as Technology (particularly AI)*.

The form in which we see this expressed is a celebration of the cleverness of humans leading to a normative flavor of congratulation and thence to an assumption of moral goodness. As with the naturalistic fallacy, troubling instances of reliance on this argument are often not clearly exposed.

The Artificialistic Fallacy addresses technology and its general endorsement of progress, and rests comfortably on ambiguous connotations of that very "progress," rather than on clear inference. Progress can mean increasing knowledge and ability, but it can also connote movement; in this case, momentum toward betterment of human life. The pragmatics of that use of the word "progress" implies that the momentum should not be stymied. Therefore, a claim of increasing technological ability can be referred to as if it were increasing human flourishing. Like the Naturalistic Fallacy (on Moore's account in terms of evolution), the Artificialistic Fallacy is dynamic, its instances seen in development over time, as trajectories. To apply the (Artificialistic) Open Question is to ask this: Given that X is an accomplishment of technology, can we still ask whether X is good? The answer to the question whether X is good could be "yes," of course, but neither possible answer begs the question.

## OVERT MANIFESTATIONS

The satirical American publication The Onion carries a recent story of Silicon Valley tech leaders promoting an automated grizzly bear as an "unavoidable and inevitable part of progress" (Onion, 2021). Making the point humorously, the piece quotes (falsely, of course) a letter from tech leaders:

> "Any kind of regulation on this front will only hinder America's ability to design and mass-produce high-quality indestructible grizzlies, which is the way the world is headed. You can't stop progress ..."

The target of this satire, of course, is the implicit claim that we *shouldn't* stop progress, no matter what.

For another example, from the Wolfram company website:

> "The rise of computation has been a major world theme for the past 50 years. Our goal is to provide the framework to let computation achieve its full poten-

> tial in the decades to come: to make it possible to compute whatever can be computed, whenever and wherever it is needed, and to make accessible the full frontiers of the computational universe." (Wolfram, 2019)

The fifty-year rise of computation constitutes the "is." The drive to compute "whatever can be computed" and make that computation accessible—its perpetuation—constitutes the "ought." To apply Moore's question is to ask whether the given the 50 years of computation rising, can we still ask whether computing whatever-can-be-computed is good? Yes, that seems a coherent question to this author.

## FOCUS AND CONTEXT

The analogy with the Naturalistic Fallacy is not perfect; there is a difference in scope between the Naturalistic and the Artificialistic Fallacies. The Naturalistic version can be invoked to express any attempt to reduce ethics to some other scale. Under that view, the Artificialistic Fallacy is an instance of the Naturalistic, in the artificial realm, rather than a peer.

Some defenders of technology detect the original Naturalistic Fallacy in a popular aversion toward modern technology in which the natural is held in higher regard than the technical. The thesis here is an inversion of that. Dorato examines the use of the term or concept "natural" in ethical arguments, along with criticism of technology as "against nature" (especially nanotechnology). He maintains that such unsupported condemnation is illegitimate (Dorato, 2015). Nothing here contradicts his claim. We agree that any argument invoking nature as support should explain and justify that move.

Nor is *techno-optimism* the target of this work. As O'Mara defines it, augmenting her account with pertinent history, techno-optimism is "the belief that technology and technologists are building the future and that the rest of the world, including government, needs to catch up" (O'Mara, 2019). Such optimism, a firm belief in the future benefits to be brought by computing technology, predicts facts, whereas the theme here is the slide from facts to ethics. (We look harder at this definition below.) Many commentators have pointed out that the leaders of Silicon Valley describe their own companies' products as making the world a better place, leaving the exact effects rather vague (ibidem). Only the future will affirm or deny the factual claims. In fact, pure techno-optimism, which is morally neutral, must be carefully factored out of ethical discussions. The related activity of *techno-evangelism* tends to conflate people's standard of living with their quality of life as a means of persuasion—promoting support, adoption, and sales—rendering its moral content indeterminate (Wikipedia, 2022). Any promoter of a particular technology product, hardware, software, or other, may be motivated simply by a conviction of the superiority of that product and the

desire to share its benefits with others, rather than motivated by normative zeal.

We call for ethical probing of the sources of such conviction, but do not condemn high tech in general. It would be superfluous to list the benefits that computing, and the Internet, have brought about. This is not criticism via the Law of the Instrument ("When you have a hammer, everything looks like a nail"), but an examination of the consequences. This is not to target individual statements, not to identify villains, nor to cultivate superiority. We all, even the most well-intentioned social observers, tend to adhere to conventional wisdom. We all need reminders.

## COVERT AND COMPARABLE MANIFESTATIONS

### *1) Marketing*

Marketing claims that glorify the computerization of processes are sometimes clearly exaggerated, such as the Salesforce statement that "Digital transformation adds value to every customer interaction" (Salesforce, 2020). Some customers with experience in call centers may dispute that, although the company concerned may indeed see added value. We expect business to promote its products, and marketing to deploy many shades of innuendo to guide consumers toward a better version of the present. So commerce may not be fair ground on which to claim foul. Yet these claims drive government and even academic research. As I opened a recent issue of the Communications of the ACM, I found this: "… for all the remarkable advancements, there's a pesky reality: smart devices could still be a whole lot more intelligent—and tackle far more difficult tasks" (Greengard, 2020). The remarkable advancements constitute the "is;" meeting the challenge of making smart devices more intelligent constitutes the "ought" (implicitly). But cannot we ask, coherently, under the remarkable advancements, whether making smart devices even more intelligent is good?

### *2) Proliferation of Technology*

Government agencies and private organizations under increased workloads are sold recommender systems to help make sensitive decisions. Government agents deploy those systems to get the job done, and also—perversely, on our view—to justify those decisions. We have seen this in the case of the criminal-sentencing system COMPAS, the bias of which (in early versions) was exposed by ProPublica (Mattu et al., 2017). This is the type of product born of the momentum of tech rather than the benefits of tech. And consider the product iBorderCtrl, intended to identify people at European border crossings. The staff of ActuIA explain it with appropriate skepticism (ActuaIA, 2019), and Gallagher and Jona later note that it fails (Gallagher et

al., 2019). This suggests that the assumption of progress-as-improvement was mistaken, but that is not really the point, because that is an error in the prediction of fact, not an error in sliding from "is" to "ought." (Failure of such a product, however, subverts the "ought" on pragmatic grounds.)

High tech reaches beyond the satisfaction of needs to the creation of needs, such as instant delivery of entertainment, smart refrigerators, and constant counting of steps taken. This paper is not a sermon on marginal or silly products of high tech. But it does have something to say about what happens during their design. Often, shortcomings or triviality are viewed as challenges to overcome, generating a technical conversation, whereas this is a call for the flaws to generate a normative conversation, allowing abandonment as a possible outcome.

***3) Commission by Omission***

The artificialistic fallacy is often committed by omission of the question "whether" in favor of the question "how," that is, in the subordination of the yes-or-no decision to elaboration on the mechanisms. Greene, Hoffman, and Stark present a study of values statements published by AI institutions, comprising non-profit, corporate, and academic membership, in which they note that " ... ethical debate is largely limited to appropriate design and implementation—not whether these systems should be built in the first place" (Greene et al., 2019). In other words, Moore's Open Question is ignored. In successive AI Now reports, the authors are increasingly alarmed by this, their recommendations moving from opening up research to monitoring AI systems to regulation and governance (Crawford et al., 2016; Campolo et al., 2017; Whittaker et al., 2018). We interpret this to indict the tech business for paying no attention. Greene and colleagues further note that the emphasis is placed on fixing AI so that its full advantages can be obtained without resistance: "... edicts to do something new are framed as moral imperatives, while the possibility of not doing something is only a suggestion, if mentioned at all" (Greene et al.).

That aligns with the point here: The normative questions about AI technology are not dismissed; rather, they simply never surface.

The attitude that "we have to get it out there" and "we have to show people" and "we have to calm their fears" are all ways of skipping the Open Question.

***4)Historical Analogy***

A couple of centuries ago, a campaign emerged in the United States that:

- Aimed to improve the lives of people and groups, even those not yet involved;
- Appealed to commercial interests and to youth who desired opportunity;

Was seen as a duty aligned with divine plan, and with nature;
Assumed that those affected would buy in when they understood the advantages;
Became a pervasive notion, fulfilled in action, while never an adopted policy.

That time was the mid 19th-century and the campaign was known as *Manifest Destiny*—the drive to settle, and thereby take over, the American West. It exhibited a marked resemblance to the current enthusiasm for high tech. Proponents such as John L. O'Sullivan saw the far Pacific Ocean, the boundary given by nature, as the right and proper extension of the new and growing United States, executing a geographic form of the Naturalistic Fallacy. A Congressman opined that God designed the original States „as the great center from which civilization, religion, and liberty should radiate and radiate until the whole continent shall bask in their blessing" (Merk, 1963, p. 28). The word "should" makes that bold declaration a moral imperative, justifying a movement already underway.

It is unfair to compare modern initiatives for AI to the militant tone of O'Sullivan's writings, in which the term "Manifest Destiny" is first used; he called on racism, uni-culturalism, and crass patriotism. Merk notes that other more generous motives, such as the spreading of democracy, the sharing of prosperity, and even the preservation of local control through federalism (states' rights), were also in play. It s important to note, in an era of opinions that sweep through the masses, that opposition was vigorous as well (ibid.). A critique of Manifest Destiny, however, should be left to real experts in political, social, and historical affairs. We are interested in the *Ought-from-Is* aspect. This historical analogy illustrates nicely the dynamic aspect of this type of fallacy; it emerges as a process rather than a static goal. The modern Artificialistic Fallacy shares that aspect but appeals but to man's ingenuity and technological prowess, rather than to God's sanction.

## CONFERRAL OF VIRTUE

As Hume says of the transition from "is" to "ought" in moral commentary, "The change is imperceptible" (Hume, 1739). To borrow phrasing from Moore, the philosophers addressed the commonplace belief that the direction in which we are developing shows us the direction in which we ought to develop. We must carefully separate the prediction from the morality.

Randy Connolly comdemns the limited vision, stating that, for too long, computing has exhibited "a tendency to rely on pop-culture theories about inevitable technology-driven social change that painted an attractive and self-satisfied veneer over our work" (Connolly, 2020, pp. 57–58). To accuse

a profession of attractive and self-satisfied work involves values, not just facts, hence is normative. Let us return to O'Mara's definition of techno-optimism: "the belief that technology and technologists are building the future and that the rest of the world, including government, needs to catch up" (O'Mara, 2019). The part about "building the future" is predictive, but the part about "needs" is normative, implying, as in Connolly, the achievement of virtue.

It's significant that the Artificialistic Fallacy is rooted not in thinking but in unthinking. No one (of whom I know) claims crudely that technology is good simply because technology is here. And, of course, much of technology is good in some sense. Virtue is a heavy load for an unconsidered assumption to carry, but virtue comes along with the normative connotations of the "ought."

And, under modern circumstances that privilege Internet communication, automated data sharing, and apps that enable quick and convenient arrangements, this conferral of virtue compounds itself, as will be described below in the section "Recursive Application."

Few would deny that technology can work out badly. See Eubanks (2018) and many other commentators for accounts of harm. But scrutiny of ethical reasoning does not have to be justified by egregious damage. Tech outcomes may be good or bad, regrettable, mild, mixed, or indifferent, but fallacies should be eschewed *anyway*. We object not to selling products nor to designing new ones, but merely to the subjugation of morals to momentum. The key pitfalls of such subjugation, described below, include vulnerability induced by the novelty and insidious recursive application of the fallacy.

## VULNERABILITY TO NOVELTY

The novelty of high tech and its attendant issues precludes cautious assessment, inflicting a vulnerability for which the public is badly prepared. The current status of iBorderCtrl is not known, as the European Commission has not released reports on its deployment in four countries in 2019 (Stolton, 2020), deprecating it as a trial project. The public had no voice in the project, initially or currently. There may be cases where that is appropriate; there may be cases where a program or facility is so new that security demands secrecy, allowing no space for serious ethical consideration. But in many new programs, normative control is unknowingly or passively abdicated to private industry, as in the case of Google's Street View—the public was never asked. Because the developments are so new as to come without normative precedents, the tech world ends up determining the suitability of its own products.

## RECURSIVE APPLICATION

Flaws or failures in artificial intelligence are blamed on the AI system, not on the attempt at application, leading to refinements in the AI and further application. The Naturalistic Fallacy allows only one iteration, which can't be repeated by humans. Because the Artificialistic Fallacy does not depend on nature, but on man, we can perpetuate it, and do.

Replies to criticism of AI application shortcomings often promise new and improved AI. According to Thomas Hellstrom, "The problem of overconfidence in AI may paradoxically increase rather than decrease over time" (Hellström, 2020). The result is repeated reworking and repeated commission of the fallacy over the previous state of affairs, a momentum toward ever more complex yet dubious technology, in a closed system.

## CONCLUSION

Many high tech companies and organization have recently undertaken ethical initiatives, but they emphasize the explanation and oversight of AI products rather than their fundamental morality. The wanton application of technology, especially machine learning and artificial intelligence, to social problems and consumer propensities reveals a particular issue in normative reasoning. Certainly, technology is sometimes ineffectual or even harmful; that is not the point here. Certainly, the assumption of inevitability of technological proliferation should not be a driving force; that is not the point here. These factors only supplement the point here—that the derivation of technical virtue, of desirability, of goodness, from the current technical trajectory, is fallacious. Insofar as the tech world itself determines the suitability of digital transformation, the tech world should take this into account.

## REFERENCES

ActuIA (No author given), *iBorderCtrl: A Dangerous Misunderstanding of What AI Really Is*, ActuIA. Jan. 12, 2019; https://www.actuia.com/english/ iborderctrl-a-dangerous-misunderstanding-of-what-ai-really-is/, accessed on 29 September 2020.

A. Campolo et al., *AI Now 2017 Report*. AI Now Institute. 2017, https://ainowinstitute. org/AI%5C Now%5C 2017%5C Report.pdf.

R. Connolly, *Why Computing Belongs Within the Social Sciences,* Communications of the ACM, 63 (8), 2020, pp. 54–59.

K. Crawford, M. Whittaker. *AI Now Report*, A summary of the AI Now public symposium, hosted by the White House and New York University's Information Law Institute, July 7th, 2016; AI Now Institute; https://ainowinstitute.org/AI%5C Now% 5C - 2016%5C Report.pdf

M. Dorato, *The Naturalness of the Naturalistic Fallacy and the Ethics of Nanotechnology*, in: The Role of Technology in Science: Philosophical Perspectives, Springer, 2015, pp. 207–224.

V. Eubanks, *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*, St. Martin's Press, 2018.
R. Gallagher, L. Jona, *We Tested Europe's New Lie Detector For Travelers—and Immediately Triggered a False Positive,* The Intercept, July 26, 2019; IBorderCtrl:%20https:// theintercept.com/2019/07/ 26/europe-border-control-ai-lie-detector/, accessed on 29 September 2020.
D. Greene, A. L. Hoffmann, L. Stark, *Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning*, in: Hawaii International Conference on System Sciences (HICSS), 2019.
S. Greengard, *AI on Edge,* In: *Communications of the ACM* 63.9, Sept., 2020, pp. 18–20; doi: 10.1145/3409977.
T. Hellström, *Dangerous Over-confidence in AI That So Far Is Too Unintelligent*, English version, Dagens Nyheter, Oct. 8, 2018;.https: //www.dn.se/debatt/dangerous-over-confidence-in-ai-that-so-far-is-toounintelligent/; accessed on 27 September 2020.
D. Hume, *A Treatise of Human Nature*, 1739.
S. Mattu, J. Angwin, J. Larson, L. Kirchner, *Machine Bias*, ProPublica, May 23, 2016; https:// www.propublica.org/article/ machine-bias-risk-assessments-in-criminal-sentencing.
F. Merk, *Manifest Destiny and Mission in American History: A Reinterpretation,* Alfred A. Knopf, New York 1963.
G. E. Moore, *Principia Ethica*, Revised, Cambridge University Press, 1993 (1903).
M. O'Mara, *The Church of Techno-Optimism,* The New York Times, September 28, 2019;https://www.nytimes.com/2019/09/28/opinion/sunday/silicon-valley-techno-optimism.html
The Onion (no author given), *Tech Leaders Justify Project To Create Army Of AI-Controlled Bulletproof Grizzly Bears As Inevitable Part Of Progress,* 2/10/22; https://www.theonion.com/tech-leaders-justify-project-to-create-army-of-ai-contr-1848402815
Salesforce (no author given), *What Is Digital Transformation*, SaleForce.com, Inc. Sept. 27, 2020; https://www.salesforce.com/products/platform/ what-is-digital-transformation/
S. Stolton *MEP, Public Has a Right to Know about Commission's Lie Detector Tech*, EURACTIV.com. Apr. 1, 2020; https://www.euractiv. com/section/digital/news/mep-public-has-a-right-to-know-aboutcommissions-lie-detector-tech/
M. Whittaker et al., *AI Now Report 2018*, AI Now Institute, 2018; https://ainowinstitute.org/AI%5C Now%5C 2018%5C Report.pdf
Wikipedia, *Technology Evangelist*; https://en.wikipedia.org/w/index.php?title=Technology_evangelist&oldid=1059138275, accessed on 27 February 2022.
Wolfram Research (No author given), *About Wolfram Research*; http://www.wolfram.com/company/background.html, accessed on 26 March 2019.

ABOUT THE AUTHOR — PhD, University of Wyoming + 1000 E. University Avenue, Department 3315, Laramie, Wyoming 82071 U.S.A.

Email: hill@uwyo.edu

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Simon X. Duan

# PLATONIC COMPUTER—THE UNIVERSAL MACHINE THAT BRIDGES THE "INVERSE EXPLANATORY GAP" IN THE PHILOSOPHY OF MIND

doi: 10.37240/FiN.2022.10.zs.14

***ABSTRACT***

The scope of Platonism is extended by introducing the concept of a "Platonic computer" which is incorporated in metacomputics. The theoretical framework of metacomputics postulates that a Platonic computer exists in the realm of Forms and is made by, of, with, and from metaconsciousness. Metaconsciousness is defined as the "power to conceive, to perceive, and to be self-aware" and is the formless, contentless infinite potentiality.

Metacomputics models how metaconsciousness generates the perceived actualities including abstract entities and physical and nonphysical realities. It is postulated that this is achieved via digital computation using the Platonic computer. The introduction of a Platonic computer into the realm of Forms thus bridges the "inverse explanatory gap" and therefore solves the "inverse hard problem of consciousness" in the philosophy of mind.

**Keywords**: Platonism, Platonic computer, pancomputationalism, metacomputics, metaconsciousness, metaprocessor, metadata, metaprogram, abstract entities, physical reality, nonphysical reality.

## 1. INTRODUCTION

In philosophy of mind, the "hard problem of consciousness" was so-named by David Chalmers (Chalmers, 1995), although the issue has scholarly antecedents from considerably earlier put forward by thinkers including John Locke (Locke, 1772) and Thomas Henry Huxley (Huxley, 1868).

The hard problem of consciousness arises from taking the position of physicalism which holds that material existence is fundamental and everything is physical. From this it is evident that consciousness is a derivative of the physical brain. However, if a physicalist world view is adopted it becomes difficult to explain how neuronal activities in the brain gives rise to the first-person conscious experience of experiences such as red rather than

green, or the sound of a dog barking, the smell of rose, or taste of red wine. There is an explanatory gap (Levine, 1983) between our understanding, no matter how complete, of the neuro correlates of a conscious experiences and our subjective conscious experience.

Having realized that the hard problem of consciousness is unsolvable, the last few decades has seen an increasing number of consciousness researchers turning away from physicalism and adopting alternative philosophical positions. Many made the progressive move to dualism in the 1980's and 1990's, then on to panpsychism in 2000's, and more recently to idealism (Chalmers 2019).

Idealism holds that consciousness is the fundamental nature of reality, that is, everything is mental. From this point of view matter is a derivative of consciousness. However, if an idealistic world view is adopted then it becomes difficult to explain how consciousness gives rise to the apparently independent existence of the material world. This is encapsulated in the phrase the "inverse hard problem of consciousness" that was coined by Max Velmans to highlight this issue (Velmans, 2021).

Various attempts have been made to address the inverse hard problem of consciousness. For example, Bernardo Kastrup put forward the argument that universal consciousness is all there ultimately is, with everything else in nature being reducible to patterns of excitation of this consciousness (Kastrup, 2019). Kastrup did not, however, propose a mechanism that would explain how such patterns of excitation of universal consciousness could give rise to the perceived phenomenal physical world.

Donald Hoffman proposed "the interface theory of perception," which postulates that the objects we perceive in time and space are metaphorical icons that act as our interface to reality (Hoffman, 2010). Hoffman uses the metaphor of a desktop computer and its icons. The icons of such a desktop computer provide a functional interface so that the user does not have to deal with the underlying programming or the electronics in order to use the computer efficiently. The interface theory of perception uses a mathematical model based around conscious agents, within a fundamentally conscious universe, to support conscious realism as a description of nature. However, the causal link between universal consciousness and the icons in the interface is not well established by the proposed mathematical model.

Hence, despite various attempts to deal with this issue there still appears to be an "inverse explanatory gap" between the perceived phenomenal physical world and universal consciousness. As such, this paper attempts to bridge the inverse explanatory gap by incorporating pancomputationalism into the framework of idealism.

## 2. PANCOMPUTATIONALISM

Human intellect often relies heavily on metaphor to approach the unknown. For example, if we cannot see the elephant, we can use conceptual metaphors such as a pillar, fans, rope, wall, etc. to comprehend and describe it.

Throughout the history of science, metaphors have been used to propose and refine scientific theories and models. This has included the metaphors of light as a wave, light as particles, gas as billiard balls, electric current as flow and the atom as a planetary system. All these are examples of metaphor-based hypotheses that have been accepted into mainstream scientific thinking and theories. Nevertheless, other metaphorical models such as the plum pudding model of the atom, were discarded when they failed to explain new experimental results.

Since the second half of the 20th century, inspired by the development of computation and digital communication technologies, some computer scientists and physicists have proposed a range of new ideas of reality that describe the universe as the output of computation. In 1969, Konrad Zuse, one of the earliest pioneers of the modern computer, first suggested the idea that the entire universe was being computed on a computer (Zuse, 1969). Mirroring this idea others, such as John Wheeler, proposed the now famous remark "it-from-bit:"

> " 'It from bit' symbolizes the idea that every item of the physical world has at bottom—a very deep bottom, in most instances—an immaterial source and explanation; that which we call reality arises in the last analysis from the posing of yes–no questions and the registering of equipment-evoked responses; in short, that all things physical are information-theoretic in origin and that this is a participatory universe." (Wheeler 1990)

Computer scientist Edward Fredkin speculated that such an idea *"*... only requires one far-fetched assumption: there is this place. Other, that hosts the engine that 'runs' the physics" (Fredkin, 2005). Other scientists, who have modelled the universe as the processing output of a giant computer include Jűrgen Schmidhuber (Schmidhuber, 1997), Stephen Wolfram (Wolfram, 2002), Max Tegmark (Tegmark, 2007), Hector Zenil (Zenil, 2012), and Tommaso Bolognesi (Bolognesi, 2012).

Furthermore, quantum versions of this computational universe hypothesis have been proposed by Nobel laureate Gerard't Hooft (Hooft, 1999), David Deutsch (Deutsch, 1997), Seth Lloyd (Lloyd, 2005), Paola Zizzi (Zizzi, 2005) and Brian Whitworth (Whitworth, 2010).

Similarly, the pancomputationalist world view was popularized in its current form by philosopher Nick Bostrom who uses a type of anthropic reasoning to claim that humans are almost certainly living in a computer simulation (Bostrom, 2003).

These previous ideas and models however have not considered how such a computer, that might be capable of generating and sustaining the universe, could have come into existence. To refine such computational universe speculations into a more coherent theory the following questions need to be addressed:

— Where is the computer?
— Where does it come from?
— What is it made of?
— How was it built?
— What are its properties?
— Who or what is the programmer?

Metacomputics has been proposed by Simon X. Duan (Duan, 2018) as a theoretical framework that is able to address each of these questions.

## 3. THE THEORETICAL FRAMEWORK OF METACOMPUTICS

Metacomputics is a theoretical framework constructed on the basis of Platonism. Platonism holds that abstract entities exist objectively in the realm of Forms, which operates as the fundamental reality, as illustrated in Figure 1. Abstract entities, such as numbers, geometric shapes and abstract objects are real and perfect nonphysical forms. Whereas material objects in the physical realm are only "shadows", or pale approximations of these abstract entities in the realm of Forms. Nevertheless, material objects in the physical reality resemble their perfect abstract forms to varying degrees, as illustrated in Figure 2.

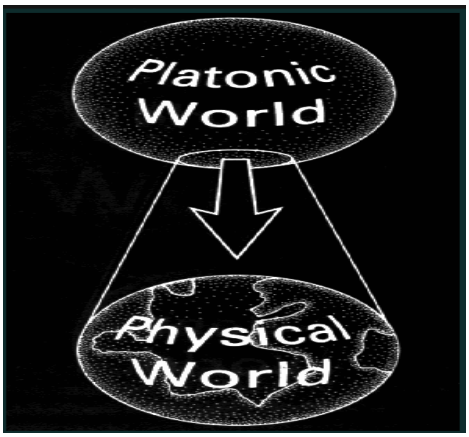



**Figure 1**. Platonic realm of Forms
Source: http://www.newforestcentre.info/realm-forms.html

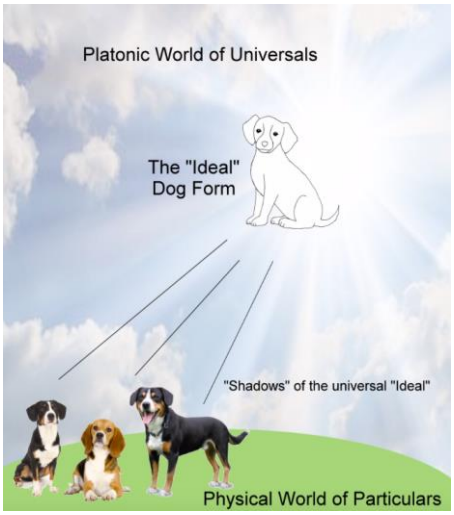



**Figure 2**. Platonic realm of Forms and physical world of particulars.
Source: https://paulspassingthoughts.files.wordpress.com/2017/02/plato-dog-form.jpg

Based on such a principle, it is postulated that the everyday material computer made of silicon is only a shadow, or poor imitation, of the perfect abstract metacomputer that exists in the realm of Forms. This abstract metacomputer is called the "Platonic computer" (Duan, 2018). The parallel existence of the physical computer and Platonic computer is illustrated in Figure 3.

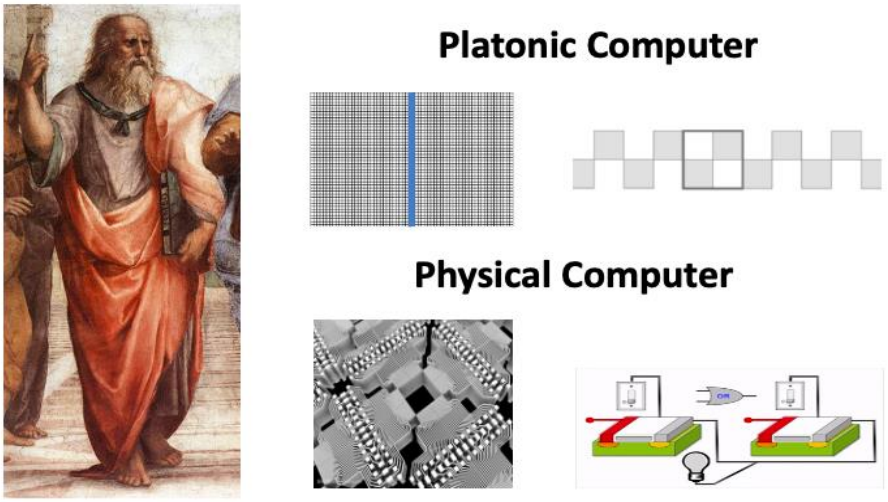



**Figure 3**. Physical computer as "shadow" or poor imitation of platonic computer

The theoretical framework of metacomputics (Duan, 2018) can be summarised by the following key points:

— The presumed existence of an operating metacomputer (i.e., Platonic computer) in the realm of Forms.
— This Platonic computer is an universal machine that is made by, of, with, and from metaconsciousness.
— Metaconsciousness is defined as the "power to conceive, to perceive, and to be self-aware" and is the formless, contentless infinite potentiality.
— Actualities arise from metaconsciousness via metacomputation of the Platonic computer.

According to metacomputics, the construction of the Platonic computer involves the following three steps (Duan, 2018), as illustrated in Figure 4:

1. Metaconsciousness manifests itself into existence.
2. Binary metaphysical switches are made with two opposing states (unmanifested metaconsciousness and manifested metaconsciousness).
3. The metacomputation system is constructed using the binary metaphysical switches to form the following three faculties: a metaprocessor, metadata, and a metaprogram.

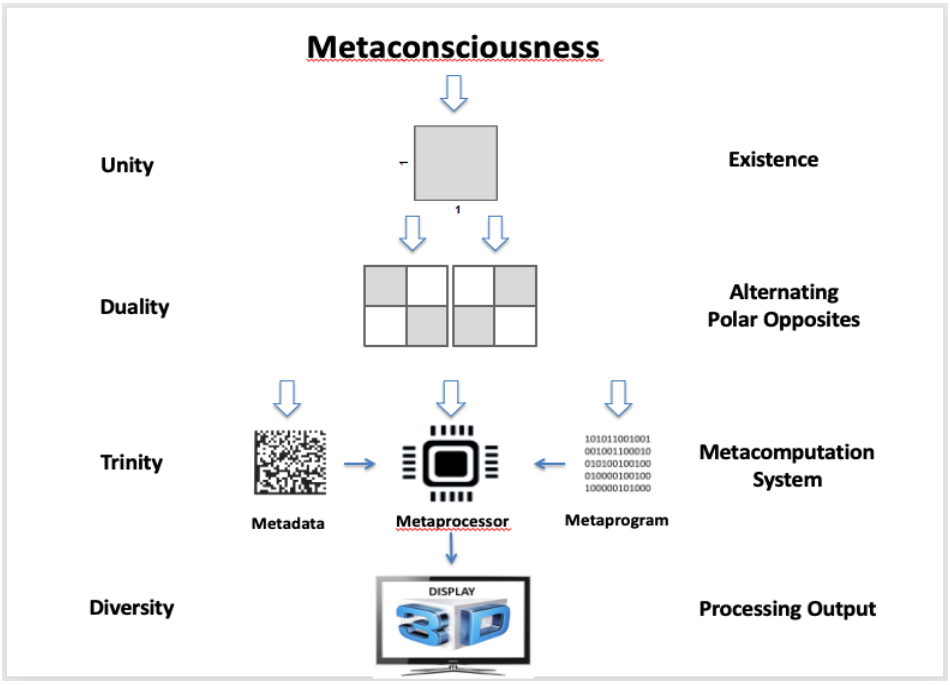



**Figure 4.** Construction of Platonic computer in the realm of Forms

The processing output of the metacomputation system gives rise to a diverse range of actualities, including abstract entities as well as physical and nonphysical realities.

## 4. COMPARISON BETWEEN PLATONIC COMPUTER AND PHYSICAL COMPUTER

Both the Platonic computer and the physical computer operate on binary opposing states. However, there are fundamental differences between the make-up of the binary states in the Platonic computer compared to the physical computer. For example, within the Platonic computer the metaprocessor is made of metaconsiousness, and the output is generated by manipulating the two binary opposing states, i.e., manifested metaconsciousness and unmanifested metaconsciousness (Duan, 2018), as illustrated in Figure 5.

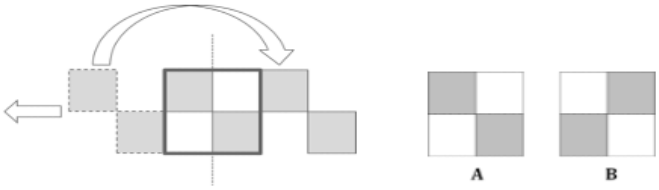



**Figure 5.** Switching of binary opposing states—Pixel (shaded square) and Void (blank square)—by alternating image A and B (Duan 2018). Pixels denote manifested metaconsciousness, voids denote unmanifested metaconsciousness

Similar to the metaprocessor, the other two faculties in the metacomputation system, i.e., metadata and metaprogram are also composed of manifested metaconsciousness (i.e., pixels or 1s) and unmanifested metaconsciousness (i.e., voids or 0s), as shown in Figure 6.

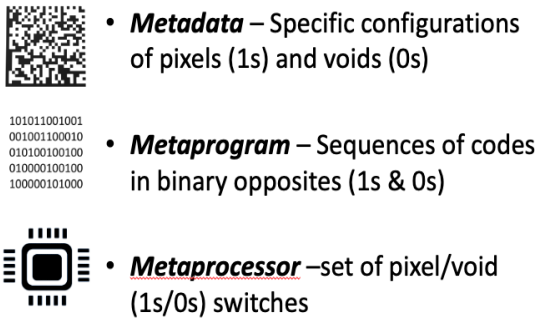



**Figure 6**. Three faculties of Platonic computer (metacomputation system) are all composed of binary opposing metaphysical states

The processing outputs of the Platonic computer are specific configurations of the binary states (i.e., manifested metaconsciousness and unmanifested metaconsciousness). Furthermore, a specific configuration of these binary states defines the conscious state of being a specific actuality, such as redness, twoness.

In comparison, a physical computer processor is made of binary ON/OFF switches made of silicon, and its output is generated by manipulating these physical switches. Hence, the processing outputs of the physical computer are specific configurations of binary states, i.e., ON (1) and OFF (0). A specific configuration of these 1s and 0s defines a symbol which can be displayed on the computer screen. For example, according to the American Standard Code for Information Interchange (ASCII), the digits 1010 defines the symbol "10," whereas the binary digits 01000001 defines the letter "A," all of which can be displayed on the computer screen.

Representing a shadow of Platonic computation, physical computation using a material computer can only simulate certain aspects of it. For instance, a physical computer can simulate the dynamic changes of the weather so that a weather forecast can be made with a reasonable level of accuracy, but it doesn't get wet and windy inside the computer screen that displays this simulation. That is, simulation of the weather in a physical computer will not produce the conscious experience of wet and windy.

Despite the limitations of physical computation, the physical computer still represents a useful tool to simulate some aspects of Platonic computation. The following section discusses how the generation of abstract entities by the Platonic computer can be simulated using a physical computer.

## 5. GENERATION OF ABSTRACT ENTITIES

1) Natural numbers

According to Platonism, numbers exist beyond space and time in the realm of Forms. That is, they are neither the cause nor the result of anything physical. Mathematicians typically describe numbers by their effects or properties (e.g., set theory). They take natural numbers as a given and accept their origin and ontology as a mystery. For instance, the German mathematician and logician Leopold Kronecker (1823–1891) is reported to have said, "God created the natural numbers; all the rest is the work of man" (Gray, 2008).

The creation of natural numbers can be simulated on a physical computer by running a +1 program, as shown in Figure 7. Running this program on a physical computer produces the output: 1, 2, 3, 4, 5, 6..... and so on as a sequence of digits on the physical computer screen. Based on the assump-

tion that the physical computer is but an imitation of the Platonic computer, it is postulated that, in parallel to the physical computer, the Platonic computer running the + 1 program generates oneness, twoness, threeness, fourness, fiveness, sixness ... and so on in the realm of Forms as specific conscious states.

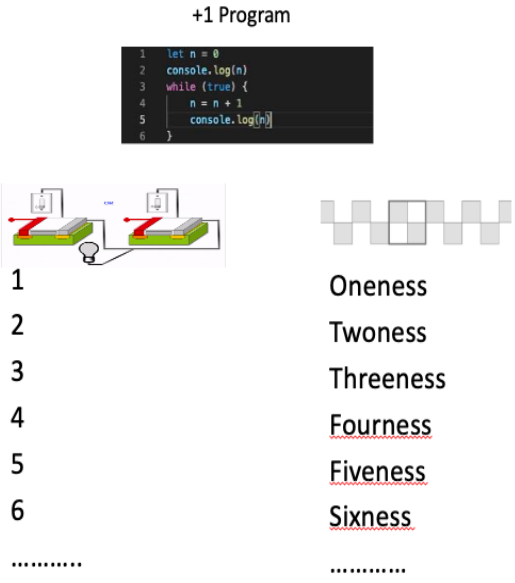



**Figure 7.** Output of a physical computer (left) and the Platonic computer (right) from running a +1 program

2) Colours

According to Platonism, colours are Universals existing beyond time and space in the realm of Forms. For example, an apple and ruby may both be red and the redness they share is a Universal and exists independently of any red "physical thing." Nevertheless, a physical computer can be programmed to produce phenomenal colours as a simulation of how the Platonic computer would produce abstract colours as Universals. As shown in Figure 8, the physical computer (on the left) generates the phenomenal colours red, green, blue. It is postulated that, in parallel to the physical computer, the Platonic computer (on the right) generates Redness, Greenness, Blueness as specific conscious states.

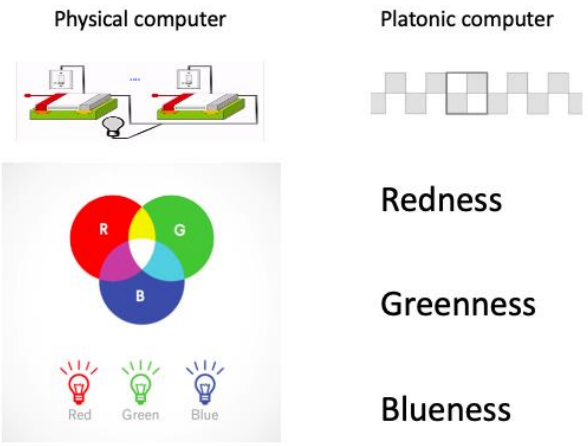



**Figure 8.** Phenomenal colours as output of a physical computer (left) and abstract colours as output of the Platonic computer (right)

## 6. GENERATION OF PARALLEL UNIVERSES

According to metacomputics, 3D space is not a given, instead, it is a processing output of metacomputation (Duan, 2018). Three-dimensional space functions as a 3D display made of voxels, as illustrated in Figure 9.

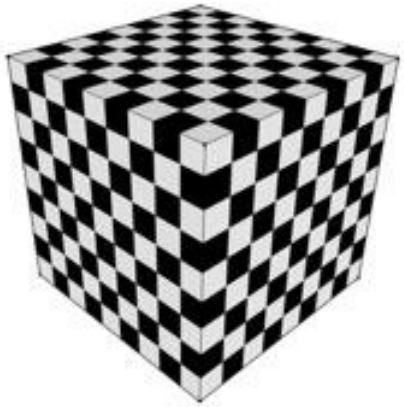


**Figure 9**. 3D space functions as a 3D display made of voxels

The contents of 3D space arise from metaconsciousness via computational processing of the Platonic computer. The Platonic computer can be configured to run at different clock speeds. At each clock speed, a universe with a specific refresh rate is displayed in the 3D space. Thus, a set of multiple parallel universes, the multiverse, is produced and sustained by metacomputation.

According to metacomputics (Duan, 2018), the memory and processing speed of the metacomputation system increases exponentially as it descends from a higher level to a lower level of parallel universe, as illustrated in Figure 10. The memory and clock speed of the Platonic computer is defined by the following formulas:

$$Memory = 4^N$$

$$Clock\ speed = 2^{(N-1)}\ cycles/unit$$

Where N is the number of levels descended from the top level of parallel universe.

With increased memory and processing capacity the Platonic computer at a lower level parallel universe would have the capacity to run more complex programs and process more complex data structures.

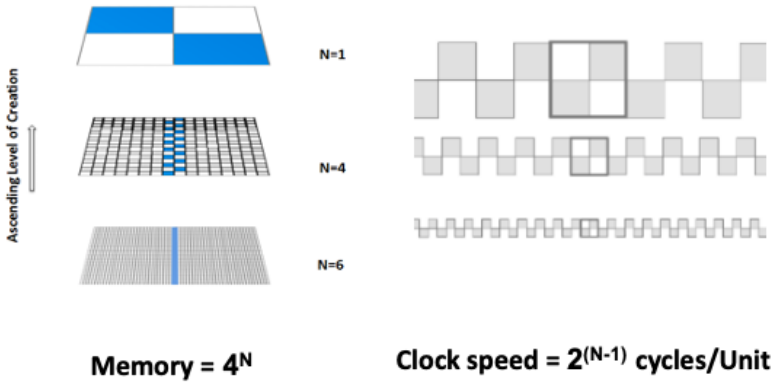



**Figure 10**. Memory and clock speed of metacomputer for different levels of parallel universes

Metacomputics postulates that each parallel universe is the representation of processing output of metacomputation by a Platonic computer operating at different clock speeds. Hence, the physical universe is only one of many parallel universes in the multiverse.

The physical universe we normally experience in our waking state is governed by a set of rules that gives rise to a set of programs called "physics" that makes the processing output display "physical." Laws of nature expressed by mathematical equations in physics are the approximation of metaprograms that govern the physical universe. For example, a falling apple is programmed to fall with an acceleration of 9.8m/s$^2$. Magnets are programmed to attract or repel each other dependent upon their polarity. In quantum mechanics, we can model the entanglement of two particles, but we cannot explain why entanglement occurs. From the point of view of metacomputics, the two particles are essentially two images on a 3D display. As they are the processing output of metacomputation, they can be programmed to entangle.

We experience physical existence such as solidity and stability because physical objects are programmed to exhibit these physical properties. As such, the physical properties of matter are derived from their programming. This can, for example, be simulated by a computer game of pool. The pool balls are images displayed on the screen that are programmed to behave as though they are solid balls.

Other parallel universes constructed by the Platonic computer operating at different clock speeds may have different rules and programs other than "physics," so they do not appear to be physical, but it does not mean they are less real. For example, those who have had a profound nonphysical visual experience during a lucid dream, or during meditation, or when using psychedelics, or when near death, etc., insist that their experience of the nonphysical realm is somehow realer than what they experience in physical realm.

As different parallel universes are processing outputs of metacomputation displayed in 3D space at different refresh rates, a specific refresh rate of the display gives rise to a specific vibration frequency of that universe. Thus, different parallel universes can potentially be experienced by tuning into their specific vibration frequencies. This can be likened to delete spaces tuning into different radio or television channels.

Although all parallel universes superimpose in the same 3D space, most people in a normal waking state tune into only one vibration frequency and perceive only the physical universe. Nevertheless, experiencing nonphysical reality is not as rare as most of us may assume. For instance, everyone experiences nonphysical visions in a dream state, many people also experience nonphysical visions in a psychedelic state, out-of-body state, meditation state and trance state, etc. Furthermore, some skilled meditation practitioners can tune into more than one vibration frequency in a controlled manner; hence they can perceive more than one parallel universe through their mind vision.

## 7. INDIVIDUAL CONSCIOUSNESS AND PERSONAL SUBJECTIVITY

Metaconsciousness is defined as the potential power to conceive, perceive and to be self-aware (Duan, 2018). According to metacomputics, metaconsciousness is the most fundamental aspect of existence that creates the Platonic computer as a universal machine and runs the metacomputation. Metaconsciousness is also the source of subjectivity, the self, the I-ness. As metaconsciousness is the creator of time and space, it transcends both time and space. Thus, metaconsciousness is nonlocal—it has no location in space, it is neither here nor there. It is also timeless—it does not come into existence or go out of existence. In addition, it is formless—it has no boundary, shape or size.

In order to experience creation in space and in time, metaconsciousness constructs multiple living beings in the multiverse via metacomputation. Metaconsciousness then fragments itself into multiple individual conscious agents and localises each individual consciousness within each living being. This gives rise to the sense of individual subjectivity.

As a living being exists in space and time, each individual conscious agent perceives the universe from the point of view of the individual living being. This allows metaconsciousness to experience the diversity of its creation through all the different first-person perspectives of all living beings in space and time in the multiverse.

The relationship between metaconsciousness and an individual conscious agent can be likened to the television broadcaster in the central control room and multiple cameramen at an event such as a football match. During the match different camera men are positioned at distinct locations around the football pitch and film the game from a particular point of view. Whereas, the broadcaster in the TV station control room has access to all the images gathered from all of the cameras across the multiple points of view around the football pitch.

## 8. COMPARISON BETWEEN PARADIGMS OF PHYSICAL SCIENCE AND METACOMPUTICS

The theoretical framework of metacomputics provides an alternative paradigm to the existing physical science approach. The two paradigms share the same empirical methodology. That is, both paradigms rely on experience to explore phenomenal reality both in the physical universe and in nonphysical parallel universes. The two paradigms differ, however, in ontology, epistemology and use of language.

Ontologically, the existing physical science paradigm assumes that matter is the primary form of existence and the ultimate reality. In contrast, the metacomputics paradigm assumes metaconsciousness is primary and is the ultimate reality. On epistemology, the existing physical science paradigm adopts a reductionist approach. This approach leads to an important consequence: it assumes that consciousness and its rich phenomenology is "nothing but" the set of neuronal interactions within the brain. In contrast, the metacomputics paradigm adopts a more holistic approach. Here, metaconsciousness digitalises itself to construct a universal machine—the Platonic computer. The multiverse is the processing output of such a universal machine.

With regards to language, the existing physical science paradigm uses mathematical equations to express the laws of nature and as such mathematics is the language of physical science. Whereas the metacomputics paradigm uses algorithms and data structures to express the diversity of the

multiverse. Hence, metaprogram and metadata are the language of metacomputics.

In summary, it has been shown that metacomputics is a new theoretical framework that provides an alternative perspective to reality. It helps to clarify many important concepts that have perplexed humans for millennia, including metaconsciousness, creation, the multiverse, reality, and individual consciousness. In addition, metacomputics has shown to be capable of explaining phenomena and patterns observed in the physical universe. Where, for example, the laws of nature revealed by physics may be viewed as mathematical expressions of metaprograms. Furthermore, metacomputics is also capable of explaining nonphysical reality, as perceived in parallel universes, in dream states, near death states, psychedelic states, and meditation states. Such perceptions are the result of the processing output of metacomputations across parallel universes.

## 9. TESTABLE PREDICATIONS OF METACOMPUTICS

As we have seen the proposed theoretical framework of metacomputics can be used to explain perceived realities, both physical as well as nonphysical. However, to qualify as a scientific theory the explanation of existing phenomena alone is not enough. A candidate scientific theory is required to provide testable predictions so that the validity of the hypotheses can be falsified or verified by experiments.

According to metacomputics, all parallel universes are the processing output of the Platonic computer. From this it is proposed that 3D space is not a given, instead, it is constructed and sustained by metacomputation. It functions as a 3D display made of voxels. This leads to the hypothesis that space itself is discreet, instead of continuous.

Zeno's paradox, commonly referred to as "The Dichotomy," is visually outlined in Figure 11 and used as a thought experiment to test this hypothesis.

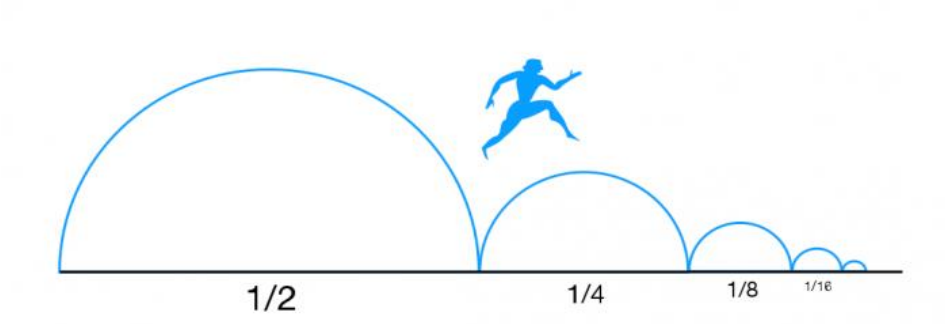



**Figure 11**. Zeno's paradox "The Dichotomy" as a thought experiment to prove space is discrete.
Source: https://commons.wikimedia.org/wiki/File:Zeno_Dichotomy_Paradox.png.
Artist: Martin Grandjean

Suppose you are going towards a point 1 m away at a rate of 1 m/s. If so, with your first step it will take you ½ second to cover half the distance, or ½ m. Let's define each additional "step" as covering half of the remaining distance in half of the remaining time. So, the 2nd step will mean covering ¼ m in ¼ second, and so on. The entire process of reaching your final destination is thus broken into infinitely many steps of ½ m, ¼ m, 1/8 m..... and so on in distance and many steps of ½ s, ¼ s, 1/8 s.....and so on in time.

$$\sum_{n=1}^{\infty}(1/2)^n = 1/2 + 1/4 + 1/8 + 1/16 + \ldots$$

Although mathematically the sum of all the steps equals 1, the number of steps is still infinity. That is, there is no "last step." So how can a process without a last step be completed?

The fact that you do reach the destination is proof that there is a last step. Therefore, there must be a smallest unit of length in space and time as the last step to reach the final point. According to metacomputics the last step is the size of the smallest voxel in the 3D display, i.e., space (see Figure 9). From the above example it can be seen that metacomputics can provide predictions which can be tested by thought experiment.

## 10. DEVELOPMENT OF PANCOMPUTATIONALISM AND PLATONISM

From the previous discussion, it can be seen that the multiverse can be modelled as the processing output of Platonic computation. Hence, metacomputics has successfully developed a pancomputationalist world view into a theoretical framework by resolving the following questions raised in section 2 of this paper;

1. Where is the computer?

The universal machine (i.e., Platonic computer) is in the realm of Forms and is beyond phenomenal space, i.e., it is nonlocal.

2. Where does it come from?

The Platonic computer is made from metaconsciousness.

3. What is it made of?

The Platonic computer is made of manifested metaconsciousness and unmanifested metaconsciousness.

4. How is it built?

The metacomputation system is built by configuring manifested metaconsciousness and unmanifested metaconsciousness into three faculties – metaprocessor, metadata and metaprogram.

5. What are its properties?

The Platonic computer operates at a specific clock speed ($2^{(N-1)}$ cycles/unit) and memory ($4^N$) within each parallel universe, where N denotes the number of levels of the parallel universe.

6. Who or what is the programmer?

Metaconsciousness is the ultimate infinite potentiality and creativity. Defragmented metaconsciousness, i.e., individual conscious agents at each level of the parallel universe perform programing for the next level of the parallel universe.

The theoretical framework of metacomputics has also extended the notion of Platonism. In the traditional Platonic realm of Forms, there are only Universals which are timeless, absolute and unchanging. The theoretical framework of metacomputics introduces an operating computer as a universal machine into the realm of Forms. Hence, the new realm of Forms contains a Platonic computer which is dynamic, and its processing output, i.e., abstract entities which are unchanging.

It is not possible to study the Platonic computer empirically as it is not accessible by our normal physical senses or by using current physical instruments. Hence, it may be difficult to accept the validity of the Platonic computer. Nevertheless, if the notion of Platonism is accepted, it can be postulated that such a universal machine exists in the realm of Forms as the archetype of the physical computer. Built on the foundation of Platonism, the theoretical framework of metacomputics can be seen to be internally consistent and coherent.

## 11. SUMMARY AND CONCLUSIONS

The scope of Platonism is extended by introducing the new concept of a "Platonic computer" which is incorporated in metacomputics. The theoretical framework of metacomputics postulates that a Platonic computer exists in the realm of Forms and is made by, of, with, and from metaconsciousness. Metaconsciousness is defined as the "power to conceive, to perceive, and to be self-aware" and is the formless, contentless infinite potentiality.

According to metacomputics, the physical computer made of silicon is but a shadow or poor imitation of the Platonic computer. Thus, by programming the physical computer it is possible to simulate the operation of the Pla-

tonic computer. The processing output of a physical computer includes symbols and virtual reality, whereas the processing output of the Platonic computer includes abstract entities as well as perceived phenomenal physical and nonphysical realities.

Metacomputics models how metaconsciousness generates the perceived actualities including abstract entities and physical and nonphysical realities. It is postulated that this is achieved via digital computation using the Platonic computer. The introduction of the Platonic computer into the realm of Forms thus bridges the "inverse explanatory gap" and therefore solves the "inverse hard problem of consciousness" in the philosophy of mind.

## REFERENCES


T. Bolognesi, *Algorithmic Causal Sets for a Computational Spacetime,* in: A Computable Universe: Understanding and Exploring Nature as Computation, H. Zenil (ed.), World Scientific Publishing, 2012.

N. Bostrom, *Are You Living in a Computer Simulation?*, Philosophical Quarterly, 53 (211), 2003, pp. 243–255; doi:10.1111/1467-9213.00309.

D. Chalmers, *Facing up to the Problem of Consciousness*, Journal of Consciousness Studies, 2 (3), 1995, pp. 200–219.

____ , 2019; https://philpapers.org/archive/CHAIAT-11.pdf, p.1

D. Deutsch, *The Fabric of Reality*, Penguin Press, Allen Lane 1997.

S. X. Duan, *Digital Consciousness and Platonic Computation: Unification of Consciousness, Mind, and Matter by Metacomputics*, American Philosophical Association Newsletter, Philosophy and Computers, 17 (2), 2018, pp. 30–40; https://cdn.ymaws.com/www.apaonline.org/resource/collection/EADE8D52-8D02-4136-9A2A-729368501E43/ComputersV17n2.pdf

E. Fredkin, *A Computing Architecture for Physics*, Computing Frontiers, 2005, pp. 273–279. Ischia: ACM.

J. Gray, *Plato's Ghost: The Modernist Transformation of Mathematics*, Princeton University Press, 2008; Archived from the original on 29 March 2017 – via Google Books.

T. H. Huxley; Youmans, William Jay, *The Elements of Physiology and Hygiene: a Text-book for Educational Institutions*, D. Appleton, New York 1868.

D. D. Hoffman, (2010-09-01). "Sensory Experiences as Cryptic Symbols of a Multimodal User Interface." Activitas Nervosa Superior. 52 (3): 95–104. doi:10.1007/BF03379572. ISSN 1802–9698.

G. 't Hooft, Quantum Gravity as a Dissipative Deterministic System, Class. Quant. Grav. 16, 1999, 3263–3279; http://arxiv.org/abs/gr-qc/9903084.

B. Kastrup, *Analytic Idealism: A Consciousness-only Ontology*, Dissertation, Radboud University Nijmegen, 2019; available on PhilArchive: https://philarchive.org/archive/KASAIA-3

J. Levine, *Materialism and Qualia: the Explanatory Gap*, Pacific Philosophical Quarterly, 64, 1983, pp. 354–361.

J. Locke, *The Works of John Locke: in Three Volumes*, vol. 1, Printed for A. Churchill, and A. Manship, and sold by W. Taylor in Pater-noster-Row, London 1722.

S. Lloyd, *The Computational Universe: Quantum Gravity from Quantum Computation*, 2005; http://arxiv.org/abs/quant-ph/0501135.

J. Schmidhuber. (1997). A Computer Scientist's View of Life, the Universe, and Everything. Lecture Notes in Computer Science, pp. 201–208. Foundations of Computer Science: Potential – Theory – Cognition. C. Freksa Ed, Springer, 1997.

M. Tegmark. (2007). The Mathematical Universe. In R. Chiao (Ed.), Visions of Discovery: Shedding New Light on Physics and Cosmology. Cambridge: Cambridge Univ. Press 2007.

M. Velmans, *Is the Universe Conscious? Reflexive Monism and the Ground of Being*, in: Consciousness Unbound: Liberating Mind from the Tyranny of Materialism, E. Kelly, P. Marshall (eds.) Rowman & Littlefield, 2021, pp.175–228; https://www.researchgate.net/publication/351308047_Is_the_Universe_Conscious_Reflexive_Monism_and_the_Ground_of_Being

J. A. Wheeler, *Information, Physics, Quantum: The Search for Links*, in: Complexity, Entropy, and the Physics of Information, W. Zurek (ed.). Addison-Wesley, Redwood City, California 1990.

B. Whitworth, *Simulating Space and Time*, Prespacetime Journal, March, 1 (2), 2010.

S. Wolfram, *A New Kind of Science*. Wolfram Media, 2003.

H. Zenil (ED.), *Introducing the Computable Universe. A Computable Universe: Understanding and Exploring Nature as Computation*. World Scientific Publishing, 2012.

P. Zizzi, *Spacetime at the Planck Scale: The Quantum Computer View*, 2005; http://arxiv.org/abs/gr-qc/0304032.

K. Zuse, *Rechnender Raum* [Calculating Space], Schriften Zur Dataverarbeitung 1. PhilArchive, 1969, copy v1: https://philarchive.org/archive/ZUSRRv1

ABOUT THE AUTHOR — PhD, Metacomputics Labs, 11 St Mary Graces Court, Cartwright Street, London, E1 8NB, UK.

Email: simon.x.duan@live.com

FILOZOFIA I NAUKA
Studia filozoficzne i interdyscyplinarne
Tom 10, zeszyt specjalny, 2022

Marcin Rabiza

# DUAL-PROCESS APPROACH TO THE PROBLEM OF ARTIFICIAL INTELLIGENCE AGENCY PERCEPTION

doi: 10.37240/FiN.2022.10.zs.15

***ABSTRACT***

Thanks to advances in machine learning in recent years the ability of AI agents to act independently of human oversight, respond to their environment, and interact with other machines has significantly increased, and is one step closer to human-like performance. For this reason, we can observe contemporary researchers' efforts towards modeling agency in artificial systems. In this light, the aim of this paper is to develop a dual-process approach to the problem of AI agency perception, and to discuss possible triggers of various agency perceptions. The article discusses the agency attribution phenomenon, based on which the argument for the dual-process nature of agency perception is developed. Two distinct types of thinking (processing) involved in human reasoning on AI agency are suggested: Type 1 and Type 2. The first one is fast, automatic, routine, and often unconscious; the second is a slower, controlled, more conscious one. These two distinct types of processing can yield differing and sometimes conflicting results for human cognition and interaction. The preliminary philosophical findings may contribute to further investigations in philosophy of mind or cognitive psychology and could also be empirically tested in HCI and UX studies.

**Keywords**: artificial intelligence; perceived agency; agency attribution.

## 1. INTRODUCTION

Intentional human action has always been distinguished from machine operation, which is usually described as a repetitive and pre-programmed activity. However, if we continue to define action by the demanding features of intentions, desires, beliefs, and mental capabilities that are typical for humans (cf. Brooks, 1991), we could "miss and misunderstand the massive changes in the intelligent machine design and interactive media use that open up Pandora's box filled with thousands of agents" (Rammert, 2015, p. 62). Artificial intelligence (AI) agents differ from human ones, but they also differ from classical machines. Thanks to recent advances in machine

learning and data science, the ability of AI agents to act independently of human oversight, respond to their environment, and interact with other machines has significantly increased, and is now one step closer to human-like performance. For this reason, we can observe contemporary researchers' efforts in defining and modeling agency in artificial systems.

The rapid development of artificial intelligence technology that started at the end of the twentieth century and continues until now, has led to an explosion of various definitions of agency. According to various authors, an agent is "A system that can act on its own behalf in an environment" (Kauffman, 2000, p. 8), "A system that tries to fulfill a set of goals in a complex, dynamic environment" (Maes, 1994, p. 136), "A system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future" (Franklin, Graesser, 1996, p. 25), "Anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors" (Russell, Norvig, 1995, p. 33), "[An] embodied system [that pursues] internal or external goals by its own actions while in continuous long-term interaction with the environment in which it is situated" (Beer, 1995, p. 173), "[A] system that can initiate, sustain, and maintain an ongoing and continuous interaction with their environment as an essential part of their normal functioning" (Smithers, 1995, p. 97), etc.

One can claim, however, that such definitions are incomplete, as they strongly rely on intuitive or commonsense notions such as "acting on one's own," "pursuing one's agenda," or "being in continuous long-term interaction." As such, they leave a significant room for subjective, individual interpretation. A good definition should be able to capture the meaning of the term as used intuitively in science and everyday life, but at the same time, it should be followed by an operational and precise conceptualization of the term. In some cases, we cannot establish whether a system is a genuine agent judging solely by its behavior because others can have a different interpretation of it. Researchers try to mitigate that effect to some extent by enlisting and discussing necessary and sufficient conditions that are to be met to classify a system as a genuine agent. Conditions, that are introduced with precise, yet understandable terms.

For instance, Barandiaran, Di Paolo, and Rohde attempt to start with a simple and non-controversial description of an agent. According to them, our common, minimal understanding of agenthood relates to "A system doing something by itself according to certain goals or norms within a specific environment" (Barandiaran et al., 2009, p. 369). Then, the authors discuss three necessary conditions that are hidden within their definition: individuality, interactional asymmetry, and normativity. *Individuality* means that a system should be distinguishable from its environment. *Interactional asymmetry* refers to agents' ability to be the source of the activity

(or "modulations") towards the environment. The third condition, *normativity*, points out that the agent's interactions with the environment are not random or arbitrary, but are done following goals or norms, which provide some sort of reference conditions for this activity.

It seems there is still no consensus on what such a minimal set of conditions should comprise, and new attempts are presented frequently. Thinking about the Barandiaran et al. approach, one may argue it is impossible, or at least is arbitrary, to decouple an agent from its environment, as well as to point out a single source of asymmetrical interaction with no doubt. This reasoning is discovered, for example, in (Rammert, 2015), where the author conceptualizes machine agency as fragmented in many pieces and delegated to myriads of pro-active and cooperative sub-systems showing low-level agency themselves, that are invisible in our everyday interaction, as they are strongly coupled in linear, sequential, or aggregated ways, into opaque yet functional black-boxes. In (Rose, Jones, 2005), the authors introduce the notion of the double dance of agency, claiming that human and machine operation outcomes are not determined by either, but emergent from the process of their interaction. As interactions between humans and intelligent machines and other systems become nowadays increasingly indistinguishable, it is difficult to establish whether a system is an individual, homogenous source of activity in its environment in order to be considered a genuine agent (cf. e.g. Nass et al., 1994; Appel et al., 2012; Araujo, 2018). In human-machine networks (HMNs), agency of the individual parts may be not *prima facie* accessible, and significant cognitive work must be done to decouple relational structure, looking for individual agents in a kind of *ex-post* rationalization attempt.

Considering a range of approaches to defining AI agency, I have recently developed two meta-concepts aggregating other accounts, which are *point* and *network* notions of agency (Rabiza, 2022). By the former, I mean notions that describe conditions for agenthood related to the agent's internal, functional organization. Point notions define AI agency according to various attributes, such as those mentioned by Barandiaran et al. The latter capture agency "not as a fixed essence or a property that something or someone possesses, but as an attribute of many actors' relationships" (Rabiza, 2022, p. 3). Here, we can classify all models attempting to explain machine agency as an emergent product of a process of human-machine interaction with many intertwined and mediated "agentive participants" involved. Furthermore, I suggested that the point-network theoretical distinction follows the dual-process nature of AI agency perception in humans (Rabiza, 2022, 3). In this paper, my aim is to further develop a dual-process approach to the problem of AI agency perception, and to discuss possible triggers of various agency perceptions.

## 2. DUAL-PROCESS AGENCY PERCEPTION

Researchers tend to admit that machine agency exists, although it differs from that of humans, and is "probably the result of human interaction and perception" (Engen et al. 2016, p. 4). Taking this into account, my focus is on an epistemological perspective of the AI agency perception problem. According to conducted literature research, artificial intelligence is foremostly *perceived* (e.g. Appel et al., 2012; Araujo, 2018; Araujo et al., 2020; Banks 2019; Engen et al., 2016; Jackson, Williams, 2020; Lucas et al., 2018; Rose, Truex, 2000; Silva, 2019) and *attributed* (e.g. Ciardo et al., 2020; Förster, Althoefer, 2021; McEneaney, 2009; Moon, Nass, 1998; Rose, Jones, 2005; Zafari, Koeszegi, 2020) during a human-computer interaction (HCI).[1] This brings us to the idea that agency attribution may be related dual-process nature of AI agency perception in humans.

Various dual-process and dual-system theories have become popular in psychology and cognitive science research (cf. an overview of such theories in Frankish, 2010). Arguably, the most well-known categories in the field are Daniel Kahneman's two systems of the mind, labeled "System 1" ("automatic system") and "System 2" ("effortful system"):

> "System 1 operates automatically and quickly, with little or no effort and no sense of voluntary control. System 2 allocates attention to the effortful mental activities that demand it, including complex computations. The operations of System 2 are often associated with the subjective experience of agency, choice, and concentration." (Kahneman, 2011, p. 22)

A form of somewhat similar dual-process theory may apply to the problem of AI agency perception. I argue that there are two distinct types of thinking (processing) involved in human reasoning on AI agency: Type 1 and Type 2. The first one is fast, automatic, routine, and often unconscious; the second is a slower, controlled, more conscious one. These two distinct types of processing can yield differing and sometimes conflicting results for human cognition and interaction.

Type 1 processing triggers a point-type agency perception during HCI. Attempting to interpret intelligent machines' agent-like behavior entails making attributions about possible causal relationships and mechanisms governing their operation. In the Type 1 mode of thinking, we are likely to attribute agential status to a technical artifact thanks to a mental mecha-

[1] "Attribution" here means a process ascribing agential status (also called an "agency judgment," (Nomura et al., 2019) to artificial actors based on the perception of AI action-outcome contiguity and causality that triggers a sense of external agency. Claims on the phenomenon of attributing human-typical qualities to machines are widely supported by empirical pieces of evidence pointing to a human tendency towards both mindful and mindless anthropomorphism of artificial entities (e.g., Kim, Sundar, 2012; Nass et al., 1995; Nowak, Biocca, 2003).

nism I propose naming *the agential stance*.[2] Daniel Dennett identifies three basic mental strategies to explain the behavior of external-world objects, which he calls "stances:" the physical stance, the design stance, and the intentional stance (Dennett, 1987). The intentional stance focuses on a reason-giving explanation of an action in terms of assumed mental properties of an agent—namely that an agent has certain desires and beliefs and acts rationally towards their completion. Indeed, studies seem to indicate that humans to some extent tend to adopt an intentional stance towards artificial agents and give mentalistic explanations of their actions (Marchesi et al., 2019). The proposed agential stance can be regarded as yet more folk-psychological, instrumentally rational heuristics for predicting, explaining, and generating quasi-stable interpretations of the external world phenomena, by a mindful as well as a mindless attribution of agency (agential status assumption) to its objects[3]. Accordingly, a human user in the agential stance tends to perceive AI agents as genuine, even human-like agents, because agency attribution makes an overall HCI easier and less effortful, improving the overall first-person user experience.

One may assume Type 1 processing works spontaneously and automatically in HCIs involving AI agents and user interfaces designed for smooth social interactions, such as social robots or socially intelligent agents (SIAs) (e.g., Dautenhahn, 1998, Dautenhahn et al., 2006; Persson et al., 2001) or affective social robots (e.g. Sengers, 2002; cf. Marchesi et al., 2019). It is less likely that Type 1 thinking will occur equally in interactions with machines devoid of similar anthropomorphic cues or a human-centered design.

In those circumstances, when agency attribution towards an artificial agent is too cognitively demanding for a human user, Type 2 processing would be expected to occur. This mode of thinking triggers a different, network-type agency perception during HCI. My argument here is that, as the human mind tends to interpret and model the behavior of external-world objects for cognitive reasons, to rationalize external agency, we automatically perceive AI agents as entities entangled in stable alliances of many actors

---

[2] Agency attribution in the agential stance should not be confused with the similar category of intentional binding effect, which refers to the subjective experience of the compressed temporal interval between voluntary action and its external sensory consequence (as a result action and outcome are perceived as being closer together in time) and is sometimes used as in an implicit measure of the first-person sense of agency (SoA) (e.g., Desantis et al., 2012; Moore, Obhi, 2012; Obhi, Hall, 2011; Suzuki et al., 2019). Agential stance, on the other hand, describes a folk-psychological mechanism of attribution based on the perception of AI action-outcome contiguity and causality, triggering the sense of external action ownership (third-person SoA) in the one observing an action similar to one's own. Thus, if intentional binding is experienced in another agent's action, it might contribute to agency attribution.

[3] This approach is in line with the cognitive miser theory in psychology, claiming that a human mind has a natural tendency to avoid spending too much of a cognitive effort and simplify the thinking process whenever possible (Stanovich, 2009; 2011). As Fiske and Taylor put it: "People are limited in their capacity to process information, so they take shortcuts whenever they can" (Fiske, Taylor 1991, p. 41).

involved in the human-machine network.[4] In the agency-status analysis, a human observer looks for both a net of external relations around the agent and its internal structure. These agential powers may comprise mediated "agentive participants" such as human designers, hardware and software architectures, algorithms, datasets, and end-users with their material practices, as well as relations with other machines and human agents that are invisible in Type 1 "black-boxed" perceptions.[5] Type 2 processing is therefore a more complex and controlled way of reasoning on external action seeking the consistency and quasi-stability that comes from a coherent view of a larger "agential structure." It consists of more detailed and nuanced thought processes focused on decomposing, deconstructing, measuring, and analyzing the relational nature of HMN agency with more of a scientific or theoretic (albeit in a naïve sense) approach (cf. Crisp, Turner, 2014).

## 3. POSSIBLE TRIGGERS OF VARIOUS AGENCY PERCEPTIONS

The immediate question of "What can trigger this range of AI agency perceptions?" is vast and remains mostly unanswered in this paper. There may be many (or very few) potential factors influencing the transition between Type 1 and Type 2 thinking about AI agency, resulting in point and network types of perceptions. An empirical study could outline a possible direction for further research. Instead, I would like to pose a hypothesis about one of the potential breaking points in the perception of AI agency that is rooted in the principles of AI design.

In *Being and Time*, Martin Heidegger differentiates two phenomenological modes of tool-being that are constituted through Dasein's varying attitudes toward objects in the world: "presence-at-hand" and "readiness-at-hand" (Heidegger, 1962). As Graham Harman puts it:

> "The latter term, ready-to-hand, refers to equipment that remains concealed from view insofar as it functions effectively. Present-at-hand, the opposite term, refers to at least three different sorts of situations. In Heidegger's writings objects present in consciousness are called present-at-hand, and so are 'broken tools' that become obtrusive once they no longer function effectively, and so is the physical concept of objective matter occupying a distinct point in space-time.
> At any rate, present-at-hand and ready-to-hand are not two different types of entities. Instead, all entities oscillate between these two separate modes: the

[4] What I mean here, is that faced with no easy access to individual agency perception we tend to look for a bigger picture in order to make sense.

[5] If external action cannot be easily rationalized by assigning causative status to an object, interpretations based on network (relational, structural) characteristics become an epistemically useful mean to an end.

> cryptic withdrawal of readiness-to-hand and the explicit accessibility of presence-at-hand." (Harman, 2019, pp. 18–19)

A tool is ready-at-hand when it is perceived as a handy piece of equipment to be used for achieving some goal. It consists of multiple parts (such as a hammer comprising a head, claw, or handle) but they are usually "hidden or withdrawn realities performing their labors unnoticed" (Harman, 2019, p. 18). Most frequently, we deal with tools within this kind of practical relation, taking them for granted as items of everyday use. The moment the tool is broken, however, it becomes present-at-hand. The tool then reveals its secrets to Dasein, who now perceives it in a more scientific or theoretic perspective, concerned only with the bare factuality of its constituent parts, regardless of its usefulness, and with no subjective context involved.

*Per analogiam*, a tool malfunction can be one of many potential triggers for a dual-process AI agency perception (or a machine agency perception in general). Another potential candidate closely related to AI design patterns could be AI *interpretability*.[6] A hypothesis I would like to pose concerns a possible negative correlation between the perceived agency of AI systems and their interpretability. The more the AI system is opaque and hides its inner workings as a nontransparent and poorly interpretable black-box (while still providing smooth interaction and good user experience), the more likely a human user will adopt Type 1 thinking along with point-type agency perceptions.

On the other hand, the more AI systems implement interpretability showing their mechanisms of operation (inner workings as well as outer relations), as more explainable without the need of a mentalistic approach to generate behavioral predictions ("agential stance"), the more likely a human user will turn to Type 2 processing and network-type perceptions, trying to rationalize the role of an AI system within a more complex HMN agentive structure.

Dual-processing in AI agency perception influenced by factors such as AI interpretability may sometimes yield conflicting results. Spontaneous agency attribution in Type 1 processing may improve the overall user experience and even trigger social reactions while interacting with AI (Appel et al., 2012; Araujo, 2018; Cowley, Gahrn-Andersen, 2021; Lucas et al., 2018). Controlled and rationalized network thinking may impede agency attribution and result in "opening the black box," and objectifying AI's agential potential.

---

[6] AI is interpretable when humans can easily understand the reasoning behind predictions and decision making of the model. The more interpretable and transparent the AI agent is, the easier it is for the user to comprehend it and trust it.

## 4. CONCLUSIONS

The aim of this article was to develop a dual-process approach to AI agency perception that was previously suggested in (Rabiza, 2022), and to discuss possible triggers of various agency perceptions.

I argue that there are two distinct types of thinking (processing) involved in human reasoning on AI agency: Type 1 and Type 2. The first is fast, automatic, routine, and often unconscious; the second is slower, controlled, and more conscious. These two distinct types of processing can yield differing and sometimes conflicting results for human cognition and interaction. Type 1 processing triggers a point-type agency perception during HCI. In the Type 1 mode of thinking, we are likely to attribute agential status to a technical artifact. However, when agency attributJion towards an artificial agent is too cognitively demanding for a human user, Type 2 processing is expected to occur. This mode of thinking triggers a different, network-type agency perception during HCI. It is a more complex way of reasoning on external action, seeking the consistency and stability that comes from a coherent view of a larger "agential structure."

Using an analogy of Heidegger's broken tool analysis I propose a hypothesis on a possible negative correlation between the perceived agency of AI systems and their interpretability. The more the AI system is opaque and hides its inner workings as a nontransparent and poorly interpretable black-box, the more likely a human user will adopt Type 1 thinking along with point-type agency perception. The more AI systems implement interpretability showing their mechanisms of operation (inner workings as well as outer relations) as more explainable without the need of a mentalistic approach to generate behavioral predictions, the more likely a human user will turn to Type 2 processing and network-type perception, trying to rationalize the role of an AI system within a more complex HMN agentive structure.

The preliminary philosophical findings may contribute to further investigations in philosophy of mind or cognitive psychology and could also be empirically tested in HCI and UX studies.

## REFERENCES

J. Appel, A. von der Pütten, N.C. Krämer, J. Gratch, *Does Humanity Matter? Analyzing the Importance of Social Cues and Perceived Agency of a Computer System for the Emergence of Social Reactions during Human-Computer Interaction*, Advances in Human-Computer Interaction, 2012, pp. 1–10.

T. B. Araujo, *Living up to the Chatbot Hype: The Influence of Anthropomorphic Design Cues and Communicative Agency Framing on Conversational Agent and Company Perceptions*, Computers in Human Behavior, 85, 2018, pp. 183–189.

T. Araujo, N. Helberger, S. Kruikemeier, C. de Vreese, *In AI we Trust? Perceptions about Automated Decision-making by Artificial Intelligence*, AI & Society, 35, 2020, pp. 611–623.

A. Bandura, *Social Cognitive Theory: An Agentic Perspectiv*e, Annual Review of Psychology, 52, 2001, pp. 1–26.
J. Banks, *A Perceived Moral Agency Scale: Development and Validation of a Metric for Humans and Social Machines*, Computers in Human Behavior, 90, 2019, pp. 363–371.
K. M. Barad, *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*, 2nd ed., Duke University Press, Durham–London 2007.
X. E. Barandiaran, E. Di Paolo, M. Rohde, *Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action,* Adaptive Behavior, 17, 2009, pp. 367–386.
R. D. Beer, *A Dynamical Systems Perspective on Agent-environment Interaction*, Artificial Intelligence, 72, 1995, pp. 173– 215.
R. A. Brooks, *Intelligence without representation,* Artificial Intelligence*,* 47, 1991, pp. 139–159.
V. Chambon, N. Sidarus, P. Haggard, *From Action Intentions to Action Effects: How Does the Sense of Agency Come about?*, Frontiers in Human Neuroscience, 8, 2014, p. 320.
F. Ciardo, F. Beyer, D. De Tommaso, A. Wykowska, *Attribution of Intentional Agency towards Robots Reduces One's Own Sense of Agency,* Cognition, 194, 2020.
S. J. Cowley, R. Gahrn-Andersen, *Drones, Robots and Perceived Autonomy: Implications for Living Human Beings,* AI & Society, 2021.
R. J. Crisp, R. N. Turner, *Essential Social Psychology*, 3rd ed., SAGE Publications, Thousand Oaks, 2014.
K. Dautenhahn, *The Art of Designing Socially Intelligent Agents: Science, Fiction, and the Human in the Loop,* Applied Artificial Intelligence, 12 (7–8), 1998, pp. 573–617.
K. Dautenhahn, A. Bond, L. Cañamero, B. Edmonds, *Socially Intelligent Agents: Creating Relationships with Computers and Robots*, in: Socially Intelligent Agents. Multiagent Systems, Artificial Societies, and Simulated Organizations, vol. 3, K. Dautenhahn, A. Bond, L. Cañamero, B. Edmonds (eds.), Springer, Boston 2002, pp. 1–20.
D. C. Dennett, *The Intentional Stance*, 1st ed., MIT Press, Cambridge–London 1987.
A. Desantis, G. Hughes, F. Waszak, *Intentional Binding Is Driven by the Mere Presence of an Action and Not by Motor Prediction*, PLoS One, 7 (1), 2012.
V. Engen, J. B. Pickering, P. Walland, *Machine Agency in Human-Machine Networks; Impacts and Trust Implications*, in: Human-Computer Interaction. Novel User Experiences, Proceedings of the 18th International Conference on Human-Computer Interaction, 2016.
S. T. Fiske, S. E. Taylor, *Social Cognition: From Brain to Culture*, 2nd ed., McGraw-Hill, New York 1991.
F. Förster, K. Althoefer, *Attribution of Autonomy and Its Role in Robotic Language Acquisition,* AI & Society, 2021.
K. Frankish, *Dual-process and Dual-system Theories of Reasoning*, Philosophy Compass, 5 (10), 2010, pp. 914–926.
S. Franklin, A. Graesser, *Is It an Agent, or Just a Program? A Taxonomy for Autonomous Agents,* in: Proceedings of the Workshop on Intelligent Agents III, Agent Theories, Architectures, and Languages, Lecture notes in computer science, Springer, Berlin 1193, 1996, pp. 21–35.
R. Glanville, *Black Boxes*, Cybernetics & Human Knowing, 16, 2009, pp. 153–167.
G. Harman, *Prince of Networks: Bruno Latour and Metaphysics*, re.press, Melbourne 2009.
G. Harman, *Technology, Objects and Things in Heidegger*, Cambridge Journal of Economics, 34 (1), 2010, pp. 17–25.
M. Heidegger, *Being and Time*, J. Macquarrie, E. Robinson (trans.), Blackwell Publishing, Oxford 1962.
R. B. Jackson, T. Williams, *On Perceived Social and Moral Agency in Natural Language Capable Robot*s, in: 2019 HRI Workshop on The Dark Side of Human-Robot Interaction: Ethical Considerations and Community Guidelines for the Field of HRI, 2020.
D. Kahneman, *Thinking, Fast and Slow*, 1st ed., Farrar, Straus and Giroux, New York 2011.
S. Kauffman, *Investigations*, Oxford University Press, Oxford 2000.
Y. Kim, S. S Sundar, *Anthropomorphism of Computers: Is It Mindful or Mindless?*, Computers in Human Behaviour, 28 (1), 2012, pp. 241–250.
B. Latour, *Reassembling the Social: An Introduction to the Actor-Network Theory*, Oxford University Press, New York 2005.

R. Legaspi, Z. He, T. Toyoizumi, *Synthetic Agency: Sense of Agency in Artificial Intelligence,* Current Opinion in Behavioral Sciences, 29, 2019, pp. 84–90.
G. M. Lucas, N. Krämer, C. Peters, L. S. Taesch, J. Mell, J. Gratch, *Effects of Perceived Agency and Message Tone in Responding to a Virtual Personal Trainer,* in: Proceedings of the 18th International Conference on Intelligent Virtual Agents; Association for Computing Machinery, New York, 2018, pp. 247–254.
P. Maes, *Modelling adaptive autonomous systems*, Artificial Life, 1, 1994, pp. 135–162.
S. Marchesi, D. Ghiglino, F. Ciardo, J. Perez-Osorio, E. Baykara, A. Wykowska, *Do We Adopt the Intentional Stance Toward Humanoid Robots?*, Frontiers in Psychology, 10, 2019.
J.E. McEneaney, *Agency Attribution in Human-Computer Interaction*, in: Engineering Psychology and Cognitive Ergonomics, D. Harris (ed.), Springer, Berlin–Heidelberg 2009, pp. 81–90.
D. Mcquillan, *Data Science as Machinic Neoplatonism*, Philosophy & Technology, 31, 2018, pp. 253–272.
Y. Moon, C. Nass, *Are Computers Scapegoats? Attributions of Responsibility in Human-computer Interaction*, International Journal of Human-Computer Interaction, 49 (1), 1998, pp. 79–94.
J. W. Moore, S.S. Obhi, *Intentional Binding and the Sense of Agency: A Review,* Consciousness and Cognition, 21 (1), 2012, pp. 546–561.
A. Moreno, A. Etxeberria, *Agency in Natural and Artificial Systems*, Artificial Life, 11 (1–2), 2005, pp. 161–175.
C. I. Nass, J. Steuer, E.R. Tauber, *Computers Are Social Actors*, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York 1994, pp. 72–78.
C. I. Nass, M. Lombard, L. Henriksen, J. Steuer, *Anthropocentrism and Computers*, Behaviour & Information Technology, 14, 1995, pp. 229–238.
O. Nomura, T. Ogata, Y. Miyake, *Illusory Agency Attribution to Others Performing Actions Similar to One's Own,* Scientific Reports, 9, 2019.
K. L. Nowak, F. Biocca, *The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments,* Presence: Teleoperators and Virtual Environments, 12 (5), 2003, pp. 481–494.
S. S. Obhi, P. Hall, *Sense of Agency in Joint Action: Influence of Human and Computer Co-actors*. Experimental Brain Research, 211, 2011, pp. 663–670.
P. Persson, J. Laaksolahti, P. Lönnqvist, *Understanding Socially Intelligent Agents — A multilayered Phenomenon,* IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 31 (5), 2001, pp. 349–360.
M. Rabiza, *Point and Network Notions of Artificial Intelligence Agency,* Proceedings, 81, 2022, p. 18.
W. Rammert, *Where the Action Is: Distributed Agency between Humans, Machines, and Programs*, in: Paradoxes of Interactivity: Perspectives for Media Theory, Human-Computer Interaction, and Artistic Investigations, U. Seifert, J.H. Kim, A. Moore (eds.), transcript Verlag, Bielefeld 2015, pp. 62–91.
J. Rose, M Jones, *The Double Dance of Agency: A Socio-Theoretic Account of How Machines and Humans Interact,* Systems, Signs and Actions, 1, 2005, pp. 19–37.
J. Rose, D.P. Truex, *Machine Agency as Perceived Autonomy: An Action Perspective*, in: Proceedings of the IFIP TC9 WG9.3 International Conference on Home Oriented Informatics and Telematics: Information, Technology and Society, Kluwer, 2000, pp. 371–390.
S. J. Russell, P. Norvig, *Artificial intelligence: A modern approach*, Englewood Cliffs, Prentice Hall, New York, 1995.
P. Sengers, R. Liesendahi, W. Magar, C. Seibert, B. Muller, T. Joachims, W. Geng, P. Martensson, K. Hook, *The Enigmatics of Affect. Anonymous*, in: Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, London 2002, pp. 87–98.
J. Silva, *Increasing Perceived Agency in Human-AI Interactions: Learnings from Piloting a Voice User Interface with Drivers on Uber*, in: Ethnographic Praxis in Industry Conference Proceedings, 2019, pp. 441–456.

T. Smithers, *Are Autonomous Agents Information Processing Systems?*, in: The Artificial Life Route to Artificial Intelligence: Building Situated Embodied Agents, L. Steels, R. A. Brooks (eds.), Erlbaum, New Haven 1995.

C. Speed, M. Disley, *Intra-actions in Data-driven Systems: A Case Study in Creative Praxis*, in: Distributed Perception: Resonances and Axiologies, N. Lushetich, I. Campbell (eds.), Routledge Studies in Science, Technology and Society, Routledge, 2021, forthcoming.

K. E. Stanovich, *The Cognitive Miser and Focal Bias*, in: Rationality and the Reflective Mind, Oxford University Press, New York 2011, pp. 65–71.

___ , *The Cognitive Miser: Ways to Avoid Thinking*, in: What Intelligence Tests Miss: the Psychology of Rational Thought, Yale University Press, New Haven 2009, pp. 70–85.

K. Suzuki, P. Lush, A.K. Seth, W. Roseboom, *Intentional Binding without Intentional Action,* Psychological Science, 30 (6), 2019, pp. 842–853.

D. Swanepoel, *Does Artificial Intelligence Have Agency?,* in: The Mind-Technology Problem: Investigating Minds, Selves and 21st Century Artefacts, R. Clowes, K. Gartner, I. Hipólito (eds.), Studies in Mind and Brain, Springer, 2021, pp. 83–104.

M. Taddeo, L. Floridi, *How AI Can Be a Force for Good*, Science, 361, 2018, pp. 751–752.

M. van Rijmenam, D. Logue, *Revising the 'Science of the Organisation': Theorizing AI Agency and Actorhood*, Innovation: Organization & Management, 23, 2020, pp. 127–144.

S. Zafari, S. T. Koeszegi, *Attitudes Toward Attributed Agency: Role of Perceived Control*, International Journal of Social Robotics, 13 (5), 2020, pp. 2071–2080.

ABOUT THE AUTHOR — PhD student, Institute of Philosophy and Sociology, Polish Academy of Sciences, Nowy Świat 72, 00-330 Warsaw, Poland;

ORCID: https://orcid.org/ 0000-0001-6217-6149

Email: marcin.rabiza@gmail.com

## INFORMACJE DLA AUTORÓW

### Przygotowanie tekstów

1. Przyjmujemy teksty rozpraw i studiów do 1,5 arkusza wydawniczego (60 000 znaków ze spacjami), polemik i głosów w dyskusjach – do 0,5 arkusza (20 000 znaków ze spacjami), recenzji – do 0,4 arkusza (około 16 000 znaków ze spacjami). W uzasadnionych przypadkach dopuszczamy wyjątki. Należy je uzgodnić wcześniej z zespołem redakcyjnym.

2. Prosimy autorów o przysyłanie tekstów w edytorze Word 1997–2003, z przypisami dolnymi, a nie końcowymi.

2a. Do każdego tekstu powinno zostać dołączone streszczenie w jęz. polskim (zamieszczone na początku tekstu) oraz w jęz. angielskim (na końcu tekstu), oraz słowa kluczowe w jęz. angielskim, informacja o afiliacji autora (umieszczona pod imieniem i nazwiskiem autora).

2b. Pożądane jest dzielenie tekstu na zatytułowane rozdziały.

3. Cytowanie pozycji literatury powinno zostać przygotowane według poniższego schematu: **Monografie**: Max Scheler, *Problemy socjologii wiedzy*, przeł. Stanisław Czerniak *et al.*, PWN, Warszawa 1990, s. 32.

**Artykuły w czasopismach**: Nelson Goodman, What Should Not Be Said about Representation?, Journal of Aesthetics and Art Criticism, 1987–8, v. 46, s. 419–425.

**Rozprawy w monografiach zbiorowych**: E. Mayr, Die Darwinsche Revolution und die Wider- stände gegen die Selektionstheorie, w: J. Herbig, R. Hohlfeld (red.), Die zweite Schöpfung. Geist und Ungeist in der Biologie des 20. Jahrhunderds, Hanser, München 1990, s. 44–70.

Odsyłacze do literatury należy umieszczać na jeden ze dwóch sposobów:

A) w przypisach dolnych;

B) w zamieszonej na końcu tekstu **Bibliografii**. W takim przypadku odsyłacze do literatury powinny być umieszczone w tekście według następującego schematu: nazwisko autora, rok wydania, strony, na przykład: (Giere, 1988, s. 25).

Wybrany przez Autora sposób A) lub B) powinien być stosowany konsekwentnie w całym tekście.

C) Bibliografia winna być uporządkowana alfabetycznie, według nazwisk autorów.

4. Elementy tekstu, które Autor pragnie wyróżnić, należy pisać rozstrzelonym drukiem.

5. Tytuły i podtytuły – wypośrodkowane, półgrubą czcionką.

6. Notki (przypisy) – dolne, a nie końcowe.

7. Autorzy proszeni są przygotowanie tekstu do celów peer-blind review, czyli o niezamieszczanie w tekście informacji pozwalających zidentyfikować autora. Dane autora na pierwszej stronie tekstu zostaną usunięte przez redakcję przed przekazaniem jej recenzentom.

8. Autorzy są ponadto proszeni o ujawnienie wszystkich osób biorących udział w powstawaniu publikacji oraz ewentualnych źródeł powstawania publikacji. To rozwiązanie zastosowane przez redakcję ma zabezpieczać publikacje przed zjawiskiem ghost-writing.

9. Autorzy są też proszeni o złożenie deklaracji (także elektronicznie, w formie skanu z podpisem), że tekst przysyłany do druku nie jest przedrukiem tekstu wcześniej publikowanego.

10. Materiały należy przysłać pocztą elektroniczną na adres:

filozofiainauka@ifispan.waw.pl

11. Ewentualne diagramy, ryciny i inne formy graficzne znajdujące się w tekstach powinny być czarno-białe.

12. Wzory matematyczne powinny być zapisane w formie Word. W razie trudności możliwe są indywidualne negocjacje z redakcją.

### Proces recenzowania

Teksty nadsyłane do czasopisma są recenzowane zgodnie ze standardami peer-blind review. Szablon recenzji oraz lista recenzentów każdego wydanego tomu czasopisma jest podana na stronie internetowej czasopisma. Lista recenzentów nie jest stała. Redakcja powołuje recenzentów w zależności od tematyki przysyłanych tekstów. Daje to gwarancję oceniania tekstów przez faktycznych specjalistów problematyki rozważanej w nadsyłanych tekstach.

**Zakres tematyczny**

W czasopiśmie będzie prezentowana cała filozoficznna problematyka, która ma związki z nauką, a więc

— problematyka filozoficzna asymilująca wyniki nauki jako przedmiot swych analiz, źródła informacji lub inspiracje;
— epistemologia i metodologia;
— dociekania nad filozofią projektowaną jako nauka;
— rozważania nad relacjami pomiędzy nauką a światem życia, rzeczywistością społeczną i kulturą.

**Interdyscyplinarność**

Publikujemy także prace łączące wątki *stricte* filozoficzne z typowo naukowymi. Zamierzenie to ujmuje rozmycie i płynność granic pomiędzy nauką i filozofią.

**Multiprogramowość**

Nie wprowadzamy programowych metafilozoficznych ograniczeń. Czasopismo nie jest forum jednej tylko szkoły filozoficznej. Multiprogramowość jest promowana w czasopiśmie między innymi jako wyraz specyfiki obecnej filozofii.