FOUNDATIONS OF COMPUTING AND DECISION SCIENCES
Vol. 44 (2019) No. 1
ISSN 0867-6356
DOI: 10.2478/fcds-2019-0005 e-ISSN 2300-3405

sciendo

Mappism: formalizing classical and artificial life views on mind and consciousness

Iwo Błądek, Maciej Komosinski, Konrad Miazga *

Abstract. Throughout centuries philosophers have attempted to understand the disparity between the conscious experience and the material world – i.e., the problem of consciousness and the apparent mind–body dualism. Achievements in the fields of biology, neurology, and information science in the last century granted us more insight into processes that govern our minds. While there are still many mysteries to be solved when it comes to fully understanding the inner workings of our brains, new discoveries suggest stepping away from the metaphysical philosophy of mind, and closer to the computational viewpoint. In light of the advent of strong artificial intelligence and the development of increasingly complex artificial life models and simulations, we need a well-defined, formal theory of consciousness. In order to facilitate this, in this work we introduce *mappism*. *Mappism* is a framework in which alternative views on consciousness can be formally expressed in a uniform way, thus allowing one to analyze and compare existing theories, and enforcing the use of the language of mathematics, i.e, explicit functions and variables. Using this framework, we describe classical and artificial life approaches to consciousness.

Keywords: mind, consciousness, mathematical modeling, simulation, perception

1. Introduction

For centuries, investigating the nature of human beings has been pursued in two branches of inquiry: understanding human brain and body, and understanding human mind. While the first branch has produced some reliable knowledge on the workings of the brain, there is not so much universally accepted and uncontested knowledge regarding mind – and especially regarding the notion of *consciousness*, which is arguably the most intriguing property of mind [23].

*Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland. Email: {iwo.bladek, maciej.komosinski, konrad.miazga}@cs.put.poznan.pl.

Our original contribution to this discussion is the proposal of a simple mathematical framework called *mappism*, which can be used to unambiguously and formally describe various views on the nature of consciousness. While in this work we mainly focus on applying mappism to describe theories of mind, it can also serve as a framework in which such theories are compared, their consequences explored, and potential inconsistencies and contradictions revealed.

In this paper we distinguish two general categories of approaches to the problem of consciousness:

- *Classical approaches*, which focus on introspection and philosophical inquiry [18].
- *Artificial Life approaches*, where complex life-like and potentially conscious systems are artificially synthesized in software, hardware, or wetware [2].

We briefly characterize both categories, provide a number of illustrative examples, and represent these examples formally in *mappism*.

1.1. What is consciousness?

The question "what is consciousness?" can be reformulated as "what is the act of a subjective perception of the world and oneself?". As science and the countless optical illusions show, we perceive reality neither comprehensively nor objectively [8]. From the perspective of modern science, we can say that we evolved to perceive primarily what we needed in order to survive. This is why the visible light spectrum is limited, why we cannot hear sounds outside of a certain range of frequencies, and why the accuracy of our senses varies in different value ranges of sensory stimuli. Past experiences also significantly affect our perceptions [3, 16]. It can be thus said that the content of consciousness for each of us is the *world* as we perceive it – our representation of it, combined with the analogous representation of ourselves and our own mental processes. From the perspective of psychology, consciousness is what we have a recollection of and what we think we operate on in our lives, after it was first preprocessed by subconscious processes in the brain [12]. These high-level thoughts or perceptions we are aware of constitute what is called *conscious* mental states. In this work, we follow the view that consciousness concerns the subjective experience of one's self, while the concept of mind is more general and usually refers to the entire scheme of information processing, as implemented in a brain (or some other medium).

We do not perceive the external world as it is, and this opens the way to questioning the validity of one's perceptions. There seems to be only one undeniably true thing about the world, and it was expressed by René Descartes in his famous "cogito ergo sum" ("I think, therefore I am"). The only certain thing for human beings seems to be that they think, and by extension, that they exist in some sense. For philosophers, questions about consciousness are questions about the nature of this "I". What or who exactly knows that it thinks? What is the ontological status of the conscious thought itself? These questions are concerned with what is known in philosophy as *phenomenal consciousness* [14, 41], which focuses on the subjectiveness of experience and the perceived *phenomena* of existing objects. One of the most widely

recognized descriptions of phenomenal consciousness is Thomas Nagel's definition of consciousness as a feeling of being something [28].

The subjective perspective of a conscious observer is often overlooked by modern science, which uses the language of mathematics to objectively describe relations between physical objects. However, if we want to apply the scientific method to research phenomenal consciousness, some way of reducing that consciousness to objectively observable facts must be first developed. One of the possible methods to address this challenge is to search for the neural correlates of consciousness [19] using subjects' reports about perceived stimuli to discover the corresponding neural activity in their brains. Another method would be to design and implement a computational model that would exhibit the features (e.g., feedback loops, internal models [33]) and the behavior (e.g., utilizing memory, learning) similar to those observed in humans. Both of these methods, while not researching phenomenal consciousness directly, can grant us important insights into the way subjective perceptions change in time and how intelligent agents develop theories about their own consciousness. These methods can also advance our understanding of what it exactly means for something to be conscious, for example by providing a useful definition of consciousness based on what is observable.

Artificial life approaches that implement the second of these two methods and are described in Sect. 4, focus on the potential emergence or synthesis of consciousness in any medium (substrate), and on developing ways to define, detect, and measure it. In contrast, classical approaches to consciousness (Sect. 3) try to describe the possible ways of relating phenomenal consciousness to mind and the physical world, and to determine what kind of physical objects could be conscious.

1.2. Problems with defining consciousness

In order to separate different aspects of consciousness, David Chalmers proposed the division into the "easy" and the "hard" problem of consciousness [5]. The "easy" problem is to understand how a brain processes information. The "hard" problem is to explain the emergence of the subjective consciousness experiencing qualia, which are thought of as the most basic and indivisible subjective perceptions (e.g., color, smell). This division is now generally accepted, although some philosophers, among others Daniel Dennett, claim that it is unnecessary and that the whole concept of qualia is unjustified. In his works, Dennett points out inconsistencies in the descriptions of qualia and subjective consciousness, and calls consciousness, perhaps a little provocatively, an "illusion" [9, 10].

The fact that even the most basic notions like subjective (phenomenal) consciousness and qualia are not agreed upon makes the single, objective, and precise description of consciousness hard to achieve [38]. Reasons for this difficulty are fundamental, and were pointed out among others in the works of Ludwig Wittgenstein [39, 40]. One of the reasons is our human language, which is used both to speak about objectively measurable objects (whose properties are therefore easy to agree upon), and to speak about more abstract and subjective notions like those regarding social life. In the first

case, it is not difficult to bind the desired meaning to a language symbol (e.g., a word) by a demonstration of its usage in conjunction with pointing to some physical objects – words like "big", "red", "sky", or "chair" are learned this way. As to the second case, abstract symbols like those related to organizing social life have a meaning primarily for the members of a group (e.g., a society) that created them ("language games" [40]). Symbols such as "good", "bad", "justice", "meaning" do not express objective and easily measurable properties, and must be learned only on the basis of how they are used in sentences and what kind of behaviors they lead to. Their focus is on manipulating minds of other individuals in a society to achieve some goal, for example maintaining order, or on the realization of some cooperative effort. Having these two examples of language use in mind, consider the meaning of the word "consciousness" in the phenomenal sense. Do we learn this meaning after a demonstration of what is and what is not conscious, or should we be satisfied with some exemplary sentences attempting to elucidate what it is and how it should be talked about? If we assume that subjective consciousness cannot be currently observed and measured from the outside of each conscious agent, and that science is generally interested in objective and measurable properties, then the scientific endeavor to define and explain subjective consciousness is obviously difficult.

When we talk about consciousness in the phenomenological sense, there is nothing to point at, because phenomenological consciousness is accessible only from the first-person perspective of the subject. Yet one cannot become someone else (i.e., someone else's representation or perception of the world) and still remain the original oneself. These are some of the reasons which make it difficult to obtain an objective measurement and a verification of purely subjective experiences, as the concept of "philosophical zombies" demonstrates [4]. This problem is aggravated by complications due to the nature of human languages, i.e., the lack of objective and precise definitions of words and terms, and the fact that words lose (or compress) a lot of information when defining high-level concepts emerging from low-level processes. Note that problems with an objective measurement and verification of subjective experiences, and problems with using high-level fuzzy terms to describe complex phenomena, are not restricted to talking about human consciousness. Similar problems exist when we ask questions about other species (Do fish feel *pain*? [17] Are plants *happy*? Are viruses *alive*?) or other "universes" (Is a computer program *intelligent*? Are simulated agents evolving in some environment *hungry*?) [20]. Mappism addresses such problems by making the relationships between different levels of description explicit and by expressing them in the language of mathematics to a degree that is considered practical or useful.

2. Mappism and its toolbox

Although it may be argued whether the nature of reality is inherently mathematical or not, most people would agree that mathematics is successfully used to describe our universe. Our model of the universe is described by physical laws, all of which are expressed in the form of mathematical formulas that allow one to make very precise

	Different language definition	Same language definition
Different substance	Occurs in dualism, Sect. 3.2	May occur when different language definitions that have different substances are mapped to the same language definition (e.g., in dualism or when describing identical, possibly independent universes each with a different substance)
Same substance	A typical situation for only one universe/one substance, e.g., physicalism in Sect. 3.3	Identity of languages

Table 1: Different combinations of equality of language definitions and language substances for a pair of languages.

predictions about the world. Ultimately, physics suggests that the entire universe can be highly accurately modeled by defining only two things: its state and the rules that govern the changes of that state.

An analogous observation can be made for consciousness, which seems to consist of its content (a mental state comprised of qualia) and the rules that govern the changes of this content. Such rules could for example describe in a systematic way how our conscious experience changes in time depending on our perceptions, our subconscious processes, and the influence of the world. Following this observation, mappism was developed as a set of tools allowing to describe different theories of mind in concise, mathematical terms. It lets one illustrate the relationships between objects (such as a brain, a mind, a bacteria, or a mountain) and also between different theories – both visually and by using a formal, mathematical notation. Note that mappism allows one to describe a theory without the need to specify all the details of the implementation of every object and every function that is used in that theory; if authors of some theory did not provide specific details and only suggested that some objects and functions exist, the corresponding mappism description will only represent this much information.

2.1. Facts, languages, states, objects, and transformations

To facilitate the understanding of the tools available in mappism, the examples presented in this section will often concern a simple case of a road for which – at any point in time – we have the complete information about the traffic situation.

Theories in mappism are described using the following elements:

- **Fact**: the smallest, indivisible piece of information in a given representation, for example a pair (d, b) of a real-valued coordinate d (corresponding to the location on the road) and a Boolean value b at that coordinate (corresponding to whether this particular location is occupied by a car).
- **Language** L: a set of all possible facts in a given representation, e.g., $L = \{(d, b) \ : \ d \in \mathbb{R}, b \in \{False, True\}\}$, where L is the set of all possible combinations of a real-number coordinate d and a Boolean value b (i.e., all possible

facts about the state of the road).

Every language in mappism is characterized by its formal definition (a representation or a formal grammar) and an additional property of substance, where the substance is independent from the formal definition of that language. Therefore, two different languages may have the same formal definition or the same substance, or both the formal definition and the substance may be different, as illustrated in Table 1. However, to avoid often unnecessary formalisms, we usually do not explicitly mention substance when defining a language. Instead, we leave it to be deduced from the context by following the rules described in Sect. 2.3.

- **State**: a subset of facts from a certain language L – for example, $S = \{(1.2, False), (2.3, True), (7.1, True)\}$. In this particular example, S is a set containing three different facts. Such a state would give us only a limited information about the road, as it only describes three different locations. S could also be an infinite set containing facts about every position on the road.

- **Object**: a process or a structure that constitutes some entity consisting of one or more states – i.e, a function (potentially with no parameters), for example $Brain$, $Mind(t)$, $Universe(t)$, $Field(x, y, z, t)$, $Dog(owner, id)$, $FibbonacciSequence(n)$, $MultiplicationTable(a, b)$, $Weather(latitude, longitude, t)$, $Element(neutrons, protons)$. An object is represented as a set of states associated with certain values of parameters (e.g., time), or speaking in other terms, it is a function from values of the parameters to their corresponding states. All states of a given object must be expressed in the same language.

  For example, $Road(t) : \mathbb{R} \to \mathcal{P}(L)$, where $\mathcal{P}(L)$ denotes the power set of L, i.e., the object *Road* for a real-valued parameter t returns a state from the language L defined in the examples above. The interpretation of states of such an object could be the state of traffic on some road, where t defines the moment of the observation (in seconds since some specific moment, e.g., the opening of the road), d defines the distance from one end of the road, and b describes whether that specific position on the road is occupied or empty.

- **Parameter**: a variable on which the state of some object depends; usually time[1] t. A parameter takes a value from a certain domain. For example, time may be expressed as a real value interpreted as the number of seconds since some specific moment – t as the parameter of $Road(t)$ could mean time elapsed since the opening of the road.

[1] In this work we are not following any specific physical theory of time, and we do not explore or discuss its existence, the relationship with space, entropy, gravity, etc. We use the argument t merely to indicate that objects may evolve or change, but the descriptions we provide are given on a level much higher than specific physical equations. However, the definition of objects and transformations in mappism can definitely be provided in terms of precise equations if this is the goal and if such equations are known, as illustrated in Sect. 4.

- **Transformation**: a function which allows for computing the states of some object, based on the states of different objects or other states of the same object. Transformations such as *transition*, *limitscope*, *remap*, and *map* are described in more detail in Sect. 2.2.

To summarize: an **object** for some values of its **parameters** returns a **state** composed of **facts** from a certain **language**. For example, $Brain(t)$ returns the state of the biological neural network composed of neuron activation levels at time moment t, described in some language $Brain_{Language}$.

A **state** of an object can be expressed as a **transformation** of a **state** of another object. For example, $HumanBiology(t)$ can be expressed as $map(HumanChemistry(t))$ for any value of t.

2.2. Transformations

Transformations are very versatile: they return a changed version of an object, meaning that the set of facts, the language substance, or the language definition may be changed as a result of applying a transformation. The substance is a property of a language, and it is used to indicate the very nature of objects described in that language – for example, if $Brain$ is a physical object and $Mind$ is a metaphysical object (e.g., an immortal soul), then we may assume that their substances are different. On the other hand, we would say that the substance of a $Mountain$ and a $Bacteria$ is the same, as they both are physical, material objects.

We distinguish four types of transformations that are useful for describing various theories of mind and consciousness. The transformations belong to two categories:

- related to object transitions – the *transition* transformation. It specifies how states of objects change when values of parameters change, and how the states of the argument object affect the states of the resulting object. The *transition* transformation does not necessarily define new objects based on its arguments, and therefore the existence of the resulting object is not dependent on the existence of the argument object. This is easy to see when the resulting object is of a different substance, and therefore, it is not possible that this resulting object is a part of the same underlying object (e.g., the universe) as the argument object.

- related to representation – the *limitscope*, *remap*, and *map* transformations. They change the description of an object. Because the only thing that changes between the argument and the result is the representation, the substance of the resulting object must be the same as the substance of arguments of these transformations. The three representation-related transformations are compared in Table 2.

While the four transformation functions mentioned above are sufficient to process objects and their languages (see Tables 3 and 4), in order to describe basic principles

	Information loss	No information loss
Change of language	*map*	*remap*
No change of language	*map*, *limitscope*	*remap*

Table 2: Characteristics of representation-related transformations.

	Change in language definition	No change in language definition
Change in substance	*transition*	*transition*
No change in substance	*remap*, *map*	*transition*, *limitscope*, *remap*, *map*

Table 3: Transformations and their possible outcomes when processing languages. Each cell corresponds to a pair of languages (the input argument and the output result of a given transformation). This information is presented in a more extensive way in Table 4.

of theories discussed in this paper, other kinds of transformation functions may be introduced if needed.

In this work we use these four transformations as deterministic functions, but they could as well handle probability distributions, and non-deterministic functions such as one-to-many transformations (e.g., turning a limited-precision floating point number into a real number) could be considered in case they are necessary to describe some theory. In the examples and philosophical theories presented further we assume that parameters of objects are discrete (e.g., time $t \in \mathbb{N}$), which is sufficient to convey the principles behind the described theories. Should functions that represent objects involved in some theory be continuous (e.g., $t \in \mathbb{R}$), differential equations and derivatives could be used to describe such continuous dynamics. Another possibility not explored in this paper is a more extensive application of logical reasoning and inference to descriptions of theories in mappism.

	Change in language definition		No change in language definition	
	Change in subst.	No change in subst.	Change in subst.	No change in subst.
transition	✓		✓	✓
limitscope				✓
remap		✓		✓
map		✓		✓

Table 4: Transformations and their possible effects on the language of the returned objects compared to the input argument object. This table presents the same information as Table 3, but in a different form: a tick indicates that a given transformation can produce the effect specified in each column.

2.2.1. The *transition* transformation

The *transition* transformation is used to define how the state of some object changes as the values of its parameters change.

$$A(t+1) = transition(A(t)) \tag{1}$$

Example (1) defines what influences the next state of the object A; in this case, the next state of A is influenced by its current state. Since the *transition* transformation describes only how the state of some object A evolves with values of its parameters, it does abstract from the initial state of that object. Initial states of objects can be explicitly specified if needed, e.g., $A(0) = \{(0.0, True), (0.5, False), (1.0, False)\}$. The *transition* transformation can take more than one argument, and the substance of the object after transition may be different from the substances of the original argument objects.

For example, $Universe(t+1) = transition(Universe(t), Mind(t))$ indicates that the next state of the universe depends on both the current state of the universe, and the current state of some $Mind$ object.

2.2.2. The *limitscope* transformation

The *limitscope* transformation is used to define some object as a specific part (a subset of facts) of some other object.

$$A(t) = limitscope(B(t)) \implies A(t) \subset B(t) \subseteq L \tag{2}$$

In this definition, *limitscope* limits the scope of $B(t)$ (expressed in a language L) to its fragment $A(t)$. This means that for every value of a shared parameter t, the state $A(t)$ of the object A will be a subset of facts from the corresponding state $B(t)$ of the object B. The states of A and B are thus necessarily expressed in the same language L, and A has less facts than B. Moreover,

$$B(t) = \cup_x A_x(t) \implies \forall_x \; A_x(t) = limitscope_x(B(t)), \tag{3}$$

so if some object B can be fully decomposed into a number of smaller objects A_x, it implies that there exist such $limitscope_x$ transformations that allow limiting the scope of object B to their respective objects A_x. The implication (3) ensures that following (6) (discussed in more detail in Sect. 2.3), such objects B and A_x will always be of the same substance, even if the $limitscope_x$ transformations will not be explicitly present in the description of a theory. Note that although *limitscope* returns a subset of its argument, it is a function that may have a sophisticated internal logic, therefore we call this kind of a transformation *limitscope* instead of simply calling it *subset*. The *limitscope* transformation always takes exactly one state as an argument, and

the resulting state has the same substance of the language as the one associated with the argument object.

For example, $Brain_x(t) = limitscope(Universe(t))$ means that $Brain_x(t)$ is a purely physical description (that is, using the same language as the description of the entire physical universe) of the matter comprising some brain x.

2.2.3. The *remap* transformation

The *remap* transformation always preserves the entire information. It is used to change the language in which an object is described, or to reformulate the description of an object using the same language.

$$A(t) = remap(B(t)) \iff B(t) = remap^{-1}(A(t)) \tag{4}$$

In this definition, *remap* expresses every state $B(t)$ of the object B as a state $A(t)$ of the object A, without any information loss (the *remap* transformation is bijective). Changing the language (or reformulating a description in the same language) may facilitate easier interpretation of the state of some object. The *remap* transformation always takes exactly one state as an argument, and the resulting state has the same substance of the language as the one associated with the argument object.

For example, $DigitInteger(i) = remapWtoI(DigitWord(i))$ means that the same object can be represented in two different, yet equivalent, ways (e.g., '271' versus 'two-seven-one'). In this example, i could indicate the index of a digit in some multi-digit number. The inverse function would be $DigitWord(i) = remapWtoI^{-1}(DigitInteger(i))$. More complex examples include the Fourier transform that remaps a signal into the frequencies that make it up (and its inverse Fourier transform counterpart), or any kind of lossless compression (and decompression).

2.2.4. The *map* transformation

The *map* transformation always leads to some loss of information. It is used to change the language in which an object is described to another (usually higher level) language, or to reformulate the description of an object using the same language.

$$A(t) = map(B(t)) \implies \nexists_{map2}\ B(t) = map2(A(t)) \tag{5}$$

In this definition, *map* expresses every state $B(t)$ of the object B as a state $A(t)$ of the object A, with information loss (the *map* transformation is not injective). This information loss is expressed in (5) by stating that there does not exist a reverse mapping that would allow to fully reconstruct the original object once this object was mapped to some other object. The change of language (or using the same language to describe an object differently) may reflect the change in the level of abstraction:

description of a state in terms of chemical molecules is not the same as its description in terms of biological cells, even though both molecules and cells are spatial combinations of some basic physical building blocks. The *map* transformation always takes exactly one state as an argument, and the resulting state has the same substance of the language as the one associated with the argument object.

For example, the statement $BrainNeural(t) = map(BrainPhysical(t))$ means that $BrainPhysical(t)$ can be described in high-level terms as $BrainNeural(t)$, and that different physical states of the brain may lead to the same neural description. Another example of a lossy mapping is the description of movement of a large number of vibrating individual particles in a medium as a continuous pressure wave characterized only by its amplitude, frequency, and speed.

2.3. The concept of substance formalized

In philosophy, or more precisely in *ontology*, there is a notion of a *substance*, i.e., a fundamental building medium of which all objects or processes of a certain kind (e.g., physical objects, thoughts) are made of or based on. In the Western philosophical tradition, the beginning of this idea can be seen in the works of Plato, who distinguished, however without using the term "substance", three different kinds of entities: souls (conscious subjects), forms, and the visible world. In the 17th century, René Descartes presented the view that there are two fundamental substances in the world: mental substance (mind) and material substance. Since a philosophical notion of a substance is present in many classical theories of mind [7, 15, 18], in order to express the ontological aspect of these theories we formalize substances in mappism as a property of a language.

Following this perspective, an object can be thought of as fluctuations within the underlying substance, whereas a language is a tool used to describe these fluctuations (so if the substance were water, objects could be waves, and then states would be still frames of these waves). To ensure that a simple change of the representation of some object does not change its substance, we only allow the *transition* transformation to change the substance of its argument objects. Using mathematical formalisms, assuming that all f and g functions are any *limitscope*, *remap*, or *map* transformations from a given mappism description of some theory, two objects A and B are of the same substance

$$\begin{aligned} & substance(A_{Language}) = substance(B_{Language}) \\ & \qquad\qquad \Longleftrightarrow \\ \exists_{f_1,f_2,\ldots,f_m} \quad & A(\ldots) = (f_m \circ f_{m-1} \circ \ldots \circ f_1)(C(\ldots)) \wedge \\ \exists_{g_1,g_2,\ldots,g_n} \quad & B(\ldots) = (g_n \circ g_{n-1} \circ \ldots \circ g_1)(C(\ldots)), \end{aligned} \tag{6}$$

for every possible combination of values of parameters of two objects A and B. In this formula, C is another, mediating object (however, $C = A$ if $m = 0$ and $C = B$ if $n = 0$).

Since (6) states a biconditional logical connective, this relationship can be exploited in two ways: (i) to derive information about substances of languages based

on the mappism description of some theory, and (ii) to restrict the set of allowed mappism transformations used while describing a theory when language substances are known based on the very assumptions of this theory. For example, in (6) we can substitute a $Mountain$ for A and a $Bacteria$ for B. Let us assume that the description of some theory makes it impossible to transform one into the other directly (i.e., there is no such composition of transformations that – when applied to one of these objects – results in the other object), but they are part of a single $Universe$. Therefore, following (i), we know that their substance is the same, because they can both be derived (through a composition of representational transformations) from their common $Universe$ (which in this example is C). On the other hand, if we substitute $Soul$ for A and $Bacteria$ for B, and if a given theory assumes that the languages of these objects are of different substances, then following (ii), the mappism description of that theory cannot contain any object C such that both $Soul$ and $Bacteria$ can be derived from C using only representational transformations. This is because should such C exist, it would imply the existence of some transformation f or g prohibited by a given theory.

To demonstrate how (6) can be used to read mappism descriptions of theories throughout this paper, we will base another example on strong epiphenomenalism from Sect. 3.2.2. In strong epiphenomenalism, one could use approach (i) to realize that there must be two different substances assigned to $Brain$ and $Mind$. Then, given that we know that there are two different substances assigned to $Brain$ and $Mind$ (either based on (i) or because this information was provided explicitly by the theory itself), it follows (ii) that it would be erroneous to include in the description of that theory any additional (otherwise feasible) transformations that would be in conflict with these two assigned substances. An example of such a transformation would be a mapping from $Brain$ to $Mind$ – although it could be functionally identical with the *transition* from $Brain$ to $Mind$, it would indicate that both of these objects are of the same substance, which would be in conflict with strong epiphenomenalism.

2.4. Families of objects

When creating descriptions in mappism, it is often useful to consider a family of objects that share some commonality. Formally, objects in a family must be expressed in the same language (including its substance) and therefore they must all be derivable from some ancestor object through (6). If such an ancestor object is not defined explicitly in a given description of some theory, it is assumed that it implicitly exists in that description.

For example, one might want to consider all existing conscious minds as different objects. Since naming of objects in mappism, besides the requirement of capitalizing the first letter, is arbitrary, the following schemes can be used for this purpose:

- $MindOfPerson1$, $MindOfPerson2$, ...
- $Mind\_1$, $Mind\_2$, ...
- $Mind_1$, $Mind_2$, ...

The last scheme which employs subscripts is practical, because it allows to use quantifiers to define transformations for many such objects (i.e., a family) at once. For example,

$$\forall_x \; Mind_x(t+1) = transition_x(Mind_x(t)) \tag{7}$$

means that the state of every individual $Mind_x$ object depends only on the previous state of itself. Notice that *transition* is also indexed, which means that every $Mind_x$ has its own transition function and thus may evolve with time according to a different set of rules. When the domain of a variable is not specified directly in the set of equations, which is the case for x in the example above, then it requires a textual definition that accompanies the mappism equations. Such textual definitions are helpful when one wants to denote by x some complex phenomenon (e.g., an existing conscious mind) that cannot be (yet) precisely described in terms of other mappism objects.

We allow omitting the quantifier in order to increase the clarity of the description, so (7) becomes

$$Mind_x(t+1) = transition_x(Mind_x(t)), \tag{8}$$

but it still should be possible to identify what x is, either by providing a textual description such as "where x is any entity possessing a mind", or by inference from the equations.

Transformations indexed with some variable, like $transition_x$ above, are not necessarily the same for every x, and are not necessarily different for every x. If they are known to be the same for every x, then the subscript is misleading and should be removed. If they are known to be different for every x, an additional constraint should be added, for example: $\forall_{x_1, x_2 \neq x_1} \; transition_{x_1} \neq transition_{x_2}$.

When using a family of objects, sometimes there is also a need to consider a set of all objects from that family. For example, imagine that a new state of an individual mind depends on the states of all other minds (a "hive mind"). To represent this, we can use standard set operations:

$$Mind_x(t+1) = transition_x(\; \cup_y \{Mind_y(t)\} \;) \tag{9}$$

Curly brackets are necessary here – they result in the set containing individual $Mind_y$ objects; without these brackets, a single, merged mind would be created (cf. Sect. 2.1). Alternatively, we can define a new object:

$$AllMinds(t) = \cup_y \; \{Mind_y(t)\} \tag{10}$$

$$Mind_x(t+1) = transition_x(AllMinds(t)) \tag{11}$$

These two notation schemes are equivalent, the second one being a bit more explicit than the first one, and thus perhaps easier to understand for some readers. In both schemes, by y we denote any entity from the same set as x.

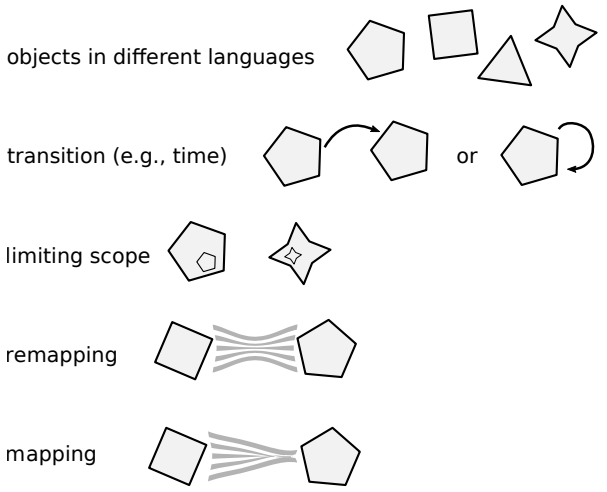

Figure 1: Visual representation of basic entities and relationships used in mappism.

2.5. Visual notation

As a shortcut and a simplification of the mathematical notation in mappism, we introduce a graphical notation where figures correspond to objects, figure shapes correspond to languages, and other symbols and arrows correspond to transformations, as shown in Fig. 1. The visual notation in most simple cases precisely reflects the mathematical notation, and in such cases both notations are equivalent. There are, however, situations where the mathematical notation is more flexible, and is therefore able to precisely describe more relationships than it would be possible in the graphical notation while keeping the picture simple and legible. Diagrams can be therefore used to convey the general idea and intuition behind some theory, and more details can be described using mathematical symbols, relations and operations like union, intersection, quantifiers, implications, etc.

2.6. Recommended usage of mappism

The minimization of the transformation inputs. In order to create concise and informative descriptions of theories, all transformations should minimize the amount of redundant information provided by arguments taken by the transformations. This means:

- Omitting redundant or unnecessary arguments.
- Using the *limitscope* transformation whenever possible to limit the scope of the object to the part of the object that contains all the necessary information.
- Using the *remap* transformation when there is no loss of information, including

cases when reformulating the description of some object is required, and when one-to-one correspondence between the elements of two languages is required.

- Using the *map* transformation to decrease the amount of information to the minimal high-level description containing all the necessary information (in order to reduce unnecessary details).
- Using the *map* transformation explicitly whenever something depends on the higher-level properties of a given object.

Preferably, the set of rules should precisely describe the system without reducing the predictability of the behavior of its objects due to the excessive reduction of information. This means that the equations should allow one to determine the probability of the next state of each described object given the probabilities of the previous states of all objects.

The use of the transition transformation. There are two conditions that should be fulfilled by transitions defined for any theory. The first condition is that the defined transitions should fully describe the evolution of all described objects as the values of their parameters change – in most cases this means the passage of time. Parameters that are nominal in nature (such as identification numbers) do not have to take part in transitions. The second condition is that redundant transitions should not be introduced unless they provide useful information (e.g., unless they constitute different and equally valid descriptions of the way some system evolves). For example, if

$$A(t+1) = transitionA(A(t)) \tag{12}$$

$$B(t) = remap(A(t)), \tag{13}$$

then it would be redundant (however still valid) to write

$$B(t+1) = transitionB(B(t)) \tag{14}$$

since we already know that

$$B(t+1) = remap(A(t+1)) = map(transitionA(A(t)). \tag{15}$$

One may also want to introduce a *transition* transformation describing an evolution of some higher-level object, such as *Brain*. It is, however, important to acknowledge that in almost all cases such a transformation would be only an approximation, as it may be impossible to precisely determine the next state of some object based only on a high-level, lossy description of its current state. In such cases, it is recommended to use the approximation symbol instead of the equality symbol to indicate that the next state of an object, computed by such a high-level *transition* transformation, may differ from the actual next state, e.g.:

$$Brain(t+1) \approx transition(Brain(t)). \tag{16}$$

Another example of such an approximated transition is the description of the laws of physics in classical (Newtonian) or relativistic (Einsteinian) mechanics. As they deal with a higher level of abstraction of a description (i.e., $UniverseNewton$ and $UniverseEinstein$) than the true reality (i.e., $UniverseTrue$), we can write the following:

$$UniverseTrue(t+1) = transitionTrue(UniverseTrue(t)) \tag{17}$$
$$UniverseEinstein(t) = mapEinstein(UniverseTrue(t)) \tag{18}$$
$$UniverseNewton(t) = mapNewton(UniverseEinstein(t)) \tag{19}$$
$$UniverseEinstein(t+1) \approx transitionEinstein(UniverseEinstein(t)) \tag{20}$$
$$UniverseNewton(t+1) \approx transitionNewton(UniverseNewton(t)). \tag{21}$$

Here, $transitionTrue$ (17) is the true, exact set of rules governing the universe (although these rules are still unknown to us). In contrast, $transitionEinstein$ (20) is only an approximation of (17), because in general

$$mapEinstein(UniverseTrue(t{+}1)) \neq transitionEinstein(mapEinstein(UniverseTrue(t))).$$

Suggested notation. Due to the fact that in mappism both objects and transformations are functions, in order to distinguish between them we recommend to use the so-called UpperCamelCase for the names of objects (e.g., $Universe$, $MindOfGeorge$), and lowerCamelCase for the names of transformations (e.g., $transition$, $mappingBrainToMind$).

The names of the specific transformations used in a description of some theory are not imposed by mappism, although the general recommendation is that the type of a transformation should be a part of the name in order to facilitate the readability of the description. Unless two transformations are exactly the same (for the same input they will always return the same output), they must have two different names.

We suggest to limit the usage of subscripts with variables (and constants) to families of objects, as described in Sect. 2.4.

2.7. Alternative ways to describe theories in mappism

It is important to keep in mind that in mappism, there are more than just one correct description of any theory. Although many descriptions may be correct (i.e., not contradictory with the theory being described), some of them may be more elegant, concise, or informative than others.

An example could be given based on the seemingly hierarchical structure of our universe. At the base level (or rather what is the base level according to our current knowledge), the state of the universe can be described as the state of its physical fields. On a higher level, field excitations can be described – with an imperfect accuracy due to, for example, the uncertainty principle – as a set of particles, each with a number of different properties. These basic particles can be later mapped to atoms and

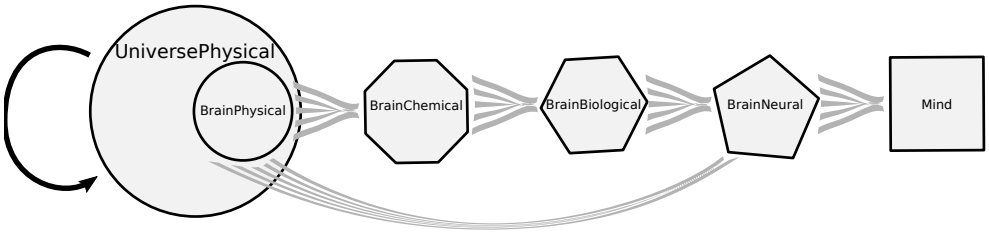

Figure 2: Visual notation of composed *map* transformations.

molecules, molecules can be mapped to cells, cells to tissues, tissues to organs such as a brain, a brain to a neural network, and – eventually – the activation levels of neurons in this network to some internal perceptions, thoughts, and emotions.

Separate sciences have been delegated to research different levels of this hierarchy, with fields and particles being in the domain of physics, molecules in the domain of chemistry, cells and organs in the domain of biology, and our thoughts in the domain of psychology. The domains of the sciences which are close to each other in this hierarchy will usually partially overlap, and so there is no clear cut between these sciences. Still, although higher-level sciences can be influenced by lower-level sciences in a myriad of ways, each of these sciences has found a number of – mostly stochastic – rules that govern the subjects of their corresponding studies and can be explained in terms of the one-level-lower science.

A path from the basic description of the universe in terms of fields to the contents of our mind can be described in mappism in the following way, as illustrated in Fig. 2:

$$\begin{aligned} UniversePhysical(t+1) &= transition(UniversePhysical(t)) \\ BrainPhysical(t) &= limitscope(UniversePhysical(t)) \\ BrainChemical(t) &= mapPtoC(BrainPhysical(t)) \\ BrainBiological(t) &= mapCtoB(BrainChemical(t)) \\ BrainNeural(t) &= mapBtoN(BrainBiological(t)) \\ Mind(t) &= mapNtoMind(BrainNeural(t)) \end{aligned}$$

The same path could be, however, shortened to the following:

$$\begin{aligned} UniversePhysical(t+1) &= transition(UniversePhysical(t)) \\ BrainPhysical(t) &= limitscope(UniversePhysical(t)) \\ BrainNeural(t) &= mapPtoN(BrainPhysical(t)) \\ Mind(t) &= mapNtoMind(BrainNeural(t)) \end{aligned}$$

In the example above we do not skip the chemical and biological levels of description – instead, we incorporate the mapping between them and other levels into the internal logic of the $mapPtoN$ transformation, such that $mapPtoN = mapBtoN \circ mapCtoB \circ mapPtoC$. This may allow for a more concise mappism description (which is beneficial if we are not interested in the intermediate levels of

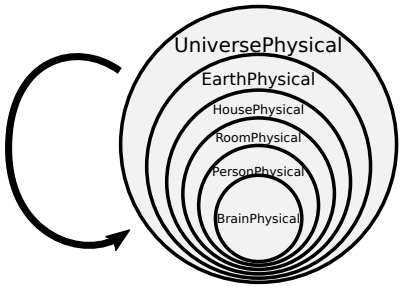

Figure 3: Visual notation of composed *limitscope* transformations.

transformations), while the earlier description could potentially be more informative, as it shows a more detailed hierarchy of abstractions built upon the physical level.

Similarly, a single *limitscope* transformation can be expanded into several *limitscope* transformations composed together, as illustrated in Fig. 3:

$$\begin{aligned} UniversePhysical(t+1) &= transition(UniversePhysical(t)) \\ BrainPhysical(t) &= limitscopeUtoB(UniversePhysical(t)) \end{aligned}$$

could be expanded to a sequence of transformations such as:

$$\begin{aligned} UniversePhysical(t+1) &= transition(UniversePhysical(t)) \\ EarthPhysical(t) &= limitscopeUtoE(UniversePhysical(t)) \\ HousePhysical(t) &= limitscopeEtoH(EarthPhysical(t)) \\ RoomPhysical(t) &= limitscopeHtoR(HousePhysical(t)) \\ PersonPhysical(t) &= limitscopeRtoP(RoomPhysical(t)) \\ BrainPhysical(t) &= limitscopePtoB(PersonPhysical(t)) \end{aligned}$$

where $limitscopeUtoB = limitscopePtoB \circ limitscopeRtoP \circ limitscopeHtoR \circ limitscopeEtoH \circ limitscopeUtoE$.

Merging and unifying descriptions on subsequent levels developed by different fields of science could therefore lead to building a "theory of everything" as the bridge from the lowest physical level to the highest one, with a huge loss of information. A far analogy and a humorous example of such an extreme information loss when going from complex low-level object description to the highest level would be the number 42 as "the Answer to the Ultimate Question of Life, The Universe, and Everything" [1], cf. [20].

3. Classical approaches to mind and consciousness

In this section, we will briefly describe the most recognized, classical theories of mind and consciousness, and we will provide their respective descriptions using mappism.

The goal here is to demonstrate how mappism can be applied to formalize the main principles of these theories, and to discuss some of the consequences of such formal descriptions. For a comprehensive treatment of the theories, we relegate the interested reader to textbooks on this topic [7, 15, 18].

Some objects and parameters commonly used in the following subsections are characterized below:

- $Brain$ – an information-processing organ composed of many relatively simple cells (neurons), which directs the behavior of humans and most animals.
- $Mind$ – a mind of an individual. As a mind, we understand here the totality of conscious thought processes, including subjective sensory perceptions.
- $Universe$ – a physical universe.
- $Mindverse$ – a "universe of minds", the union of facts describing all individual minds.
- $x, y, z, \ldots$ – entities possessing a mind. Detailed criteria for what kind of entities possess a mind are often not specified explicitly. This is a limitation of some of the theories that we describe in this section (e.g., dualism just assumes the existence of such entities).
- t – a parameter representing a time moment.

Definitions of all objects and transformations introduced in each subsection are restricted to that subsection, i.e., the $Mind$ object in interactionism (28) in Sect. 3.2.1 may be defined differently than the $Mind$ object in epiphenomenalism (35s) in Sect. 3.2.2 (despite the same name). Analogously, *limitscope* in physicalism (40) in Sect. 3.3 is not the same transformation as *limitscope* in functionalism (46) in Sect. 3.5.

3.1. Solipsism

According to solipsism, there exists only a single mind, and the perceived reality is merely imagined by that mind in a similar fashion as it happens in dreams. While other people seem to act as if they possessed minds, they are imagined as well. René Descartes's observation that the only thing one can be truly certain of is that one exists [11] can be seen as a starting point to solipsism, because the existence of anything beyond oneself can be doubted.

In mappism, solipsism can be described as follows (see also Fig. 4):

$$Mind(t+1) = transition(Mind(t)) \tag{22}$$

$$UniversePhenomenal(t) = limitscopeUniverse(Mind(t)) \tag{23}$$

$$BrainPhenomenal(t) = limitscopeBrain(UniversePhenomenal(t)) \tag{24}$$

The fundamental object in solipsism is $Mind$. A state of the $Mind$ object continually changes, and we express this by using the parameter t representing a time

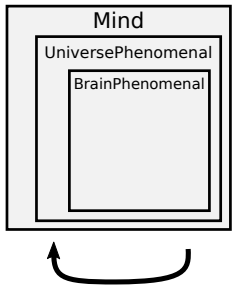

Figure 4: Illustration of solipsism using visual notation from Sect. 2.5.

moment, and by using a *transition* transformation (22) from $Mind(t)$ to $Mind(t+1)$. $UniversePhenomenal$, which is the perceived universe, is a *limitscope* of $Mind$, and thus it is the direct content of a mind (i.e., phenomena). "Phenomenal" means that this universe is only perceived in the mind, but does not exist in itself. $BrainPhenomenal$ is a *limitscope* (24) of $UniversePhenomenal$, so $BrainPhenomenal$ is only a perception of some brain, and thus does not have any causal power over $Mind$.

Since *limitscope* does not change the language, the only language used in this model is the language of $Mind$, and thus all objects are of the same substance.

3.2. Dualism

Dualism, in contrast to solipsism, assumes the existence of the material world. As mentioned in Sect. 2.3, René Descartes distinguished two fundamental substances in the world: mind and matter. This view is more widely known under the name of *Cartesian dualism*. In this view, mind is immaterial and does not follow the laws of nature; it is basically equal to consciousness, which – as Descartes believed – by its free will exerts influence on the physical body so that the body acts according to this will. The following main variants of dualism can be distinguished [15]:

- *Interactionism* – the body influences the mind with perceptual information, and the mind controls the movements of the body. This variant was originally presented by Descartes.
- *Epiphenomenalism* – the body functions on its own according to physical laws, and influences the mind with perceptual information. The mind has no control over the body – the mind is an *epiphenomenon* produced by the body.
- *Parallelism* – body and mind coexist with no real interaction.

Dualism was born long before it was understood how the brain was able to control the body. It was an attempt to explain the complexity of human behavior and human thought, when the body seemed merely like a simple automaton – a machine executing the decisions of the will. With modern science having explained the general principles of functioning of body and brain, there is only one avenue left for dualism to remain

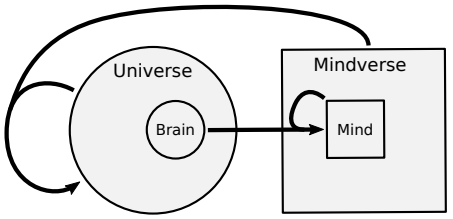

Figure 5: Illustration of interactionism.

a credible hypothesis – the still unsolved problem of consciousness and the nature of qualia.

3.2.1. Interactionism

In mappism, interactionism can be described as follows (see also Fig. 5):

$$Universe(t+1) = transitionUniverse(Universe(t), Mindverse(t)) \quad (25)$$

$$Brain_x(t) = limitscope_x(Universe(t)) \quad (26)$$

$$Mindverse(t) = \cup_x Mind_x(t) \quad (27)$$

$$Mind_x(t+1) = transitionMind_x(Mind_x(t), Brain_x(t)) \quad (28)$$

The next state of $Universe$ is influenced by its current state and the current state of all minds (25). For simplicity, the fact that each mind is limited in its influence to its corresponding brain is not represented in the above description. Each $Brain$ is simply a part of the $Universe$ (26). Each $Mind$ changes in time depending on both its current state and on the current state of the $Brain$ associated with this mind (28). $Mindverse$ is defined as a union of all facts describing each of the $Mind$ objects (27).

The above description is simplified, because it allows the $Mind$ to influence anything in the $Universe$ ("psychic powers"). To avoid this, we should limit the influence of $Mind$ to $Brain$ only. This requires three additional equations:

$$UniverseNonBrain(t) = Universe(t) \ - \ \cup_x Brain_x(t) \quad (29)$$

$$UniverseNonBrain(t+1) = transitionUniverse(Universe(t), \emptyset) \quad (30)$$

$$Brain_x(t+1) = transitionBrain_x(Universe(t), Mind_x(t)) \quad (31)$$

In these additional equations, we first define the $UniverseNonBrain$ object as everything in the universe besides brains (29). This is done in order to divide the $Universe$ object into two separate parts: the $UniverseNonBrain$, which adheres only to the rules of physics, and the $Brain_x$ objects, which can be additionally affected by their respective minds. Then, we use (30) and (31) to restate (25) in a more precise, careful way. In (30) we specify that no changes in the state of $UniverseNonBrain$ are ever caused directly by any $Mind_x$, which makes psychokinesis impossible. Finally, to complete the set of required transitions, in (31) we state that the $Brain_x$ objects are

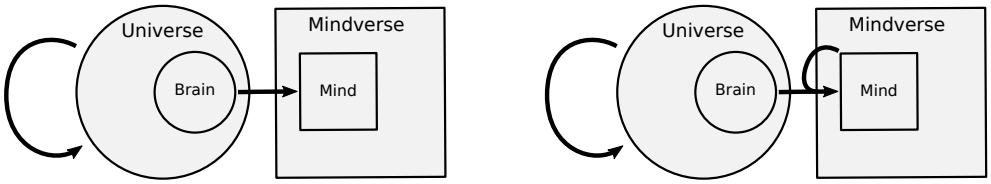

Figure 6: Illustration of epiphenomenalism. Left: strong epiphenomenalism. Right: weak epiphenomenalism.

dependent on both the $Universe$ and their respective $Mind_x$ objects. Note that (31) forbids telepathy, which previously was allowed by (25) where many $Mind_x$ objects could potentially influence the same $Brain_x$.

In this description of interactionism, there are two languages: the language of $Universe$ and the language of $Mind$. Substances of these languages are different, because their respective objects cannot be transformed to each other using representational transformations.

3.2.2. Epiphenomenalism

In mappism, epiphenomenalism can be described as follows (see also Fig. 6):

$$Universe(t+1) = transitionUniverse(Universe(t)) \tag{32}$$

$$Brain_x(t) = limitscope_x(Universe(t)) \tag{33}$$

$$Mindverse(t) = \cup_x Mind_x(t) \tag{34}$$

Strong epiphenomenalism:

$$Mind_x(t+1) = transitionMind_x(Brain_x(t)) \tag{35s}$$

Weak epiphenomenalism:

$$Mind_x(t+1) = transitionMind_x(Mind_x(t), Brain_x(t)) \tag{35w}$$

Epiphenomenalism can be expressed much more simply than interactionism. In (32), $Universe$ depends solely on the earlier state of the $Universe$, contrary to interactionism where it depended also on the state of all the existing minds. $Mind$ is influenced only by the previous state of $Brain$ in the strong variant of epiphenomenalism (35s), and by the previous states of both $Brain$ and $Mind$ in the weak variant (35w). $Mindverse$ is defined as a union of all facts describing each of the $Mind$ objects (34).

In this formal description, similarly to interactionism, there are two languages: the language of $Universe$ and the language of $Mind$. Substances of these languages are different, because their respective objects cannot be transformed to each other using representational transformations.

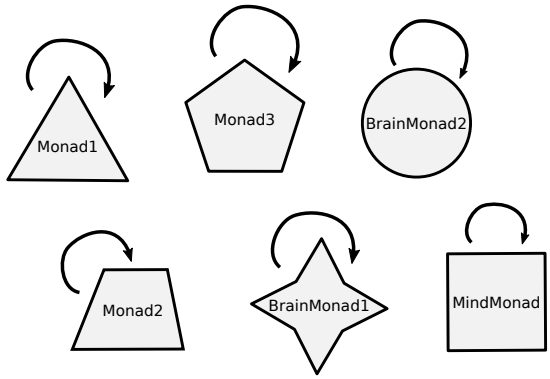

Figure 7: Illustration of parallelism in the variant of Leibniz's monadology. All objects shown are separate monads. The BrainMonad objects constitute a single brain.

3.2.3. Parallelism (Leibniz's monadology)

Parallelism is a family of views which, just as all dualistic views, assert the existence of different substances, but avoid the problem of explaining the interactions between them by claiming that there are no such interactions. This requires, however, an explanation of the apparent correlations between their behaviors. Probably the most known example of parallelism is Leibniz's monadology. According to Leibniz, the universe is composed of *monads*, which are its fundamental irreducible elements [24]. However, material objects are not the only monads – minds are monads as well. Monads do not interact with each other. The apparent physical forces we observe are not the result of interactions between monads, but are preordained by some higher-level being ("pre-established harmony") so that it appears as if there was an interaction. In fact, even our perceptions of anything are preordained.

In mappism, parallelism can be described as follows (see also Fig. 7):

$$Monads(t+1) = \{transition_m(m) \ : \ m \in Monads(t)\} \tag{36}$$

$$BrainMonads_x(t) \subset Monads(t) \tag{37}$$

$$MindMonad_x(t) \in Monads(t) \quad \wedge \quad MindMonad_x(t) \notin BrainMonads_x(t) \tag{38}$$

$Monads$ is a set of all monads in the universe. Each monad evolves in its own preordained way, so it depends only on the previous state of itself (36). $BrainMonads_x$ is a set of monads representing a physical brain associated with some entity x (37). $MindMonad_x$ is a single monad representing the mind of that entity x, however it is not a part of $BrainMonad_x$ (38). In this description of the theory, "pre-established harmony" is implicitly realized by all transitions $transition_m$ so that their results are correlated.

In parallelism, every monad is described in its own language and is subject only to its $transition$ transformation – consequently, every monad is of a different substance.

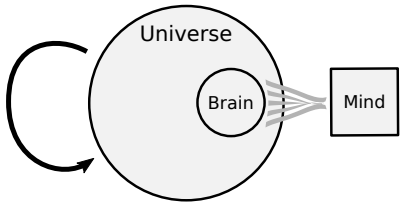

Figure 8: Illustration of reductive physicalism.

3.3. Physicalism

Physicalism is the view that everything can be explained using the language of physics, and thus that there is only one kind of substance in the world, i.e., matter. Physicalism can be divided into *reductive physicalism* and *eliminativism* (or *eliminative materialism*) [7, 18]. Reductive physicalism states that concepts like consciousness and mind will eventually be explained by purely physical processes (such concepts will be *reduced* to these processes). Currently, the most promising avenue of research in this area is the search for neural correlates of consciousness [19]. On the other hand, eliminativists claim that the concepts of consciousness and mind, as belonging to folk psychology, are false in a similar way as, for example, the concepts of ether or a life force, and thus we should stop using them in favor of something else, which is yet to be discovered.

In mappism, reductive physicalism can be described as follows (see also Fig. 8):

$$Universe(t+1) = transition(Universe(t)) \tag{39}$$

$$Brain_x(t) = limitscope_x(Universe(t)) \tag{40}$$

$$Mind_x(t) = mapMind(Brain_x(t)) \tag{41}$$

The description of physicalism in mappism is almost identical to that of the strong epiphenomenalism – the only difference is the usage of *map* instead of *transition* (41), cf. (35s). The usage of a *map* transformation means that in physicalism, $Mind$ is a high-level interpretation of the state of $Brain$, which in turn means that $Mind$ is reducible to $Brain$.

All languages in this description are of the same substance. The substance of the language of $Mind$ is the same as the substance of the language of $Universe$, because a *map* transformation is used to define $Mind_x$ objects.

3.4. Panpsychism

According to *panpsychism*, some degree of consciousness is associated with every physical process [6]. In this view, consciousness of complex objects like brain arises emergently from the consciousnesses of its subparts, or from the realized computational pattern. Since the hypothetical consciousness on the very low level of the

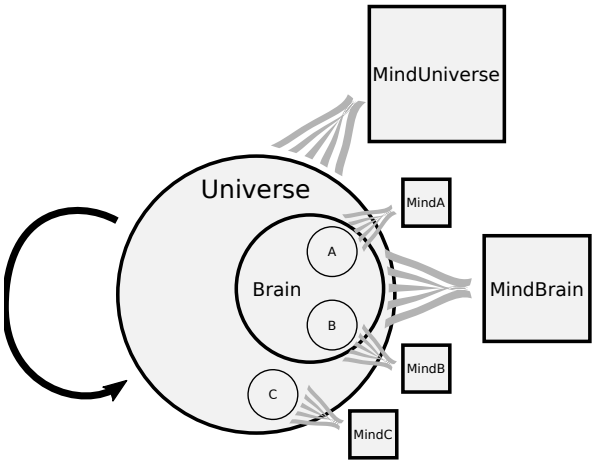

Figure 9: Illustration of panpsychism.

universe may be very simple and much different from ours, the words *psyche* and *proto-consciousness* are often used instead of the words *mind* or *consciousness*.

As for the examples of panpsychism, Erwin Schrödinger proposed that every change in the state of the universe produces some kind of (potentially extremely simple) consciousness [34]. Consciousness of complex objects, such as brains, would emerge from these low-level consciousnesses. Integrated Information Theory [36], a more recent framework created by Giulio Tononi, can also be classified as panpsychic in its nature. In this theory, the degree to which a physical system is conscious is dependent on how well it integrates information. Integrated Information Theory was later generalized by Max Tegmark for arbitrary quantum systems [35].

In mappism, panpsychism can be described as follows (see also Fig. 9):

$$Universe(t+1) = transition(Universe(t)) \tag{42}$$

$$\exists_{limitscope_s}\ Brain_s(t) = limitscope_s(Universe(t)) \tag{43}$$

$$\forall_{limitscope_s}\ Mind_s(t) = map(limitscope_s(Universe(t))) \tag{44a}$$

The main idea of panpsychism is captured in (44a), where a certain transformation assigns to every possible $limitscope_s$ of $Universe$ some subjective experience, denoted here as $Mind$ (the word "mind" is used for consistency with the descriptions of other theories). The variable s denotes a certain *limitscope* of $Universe$, and because of the universal quantifier, all possible *limitscope* transformations of $Universe$ are considered. Although the $Brain$ object is explicitly mentioned (43), it is not especially different from all the other conscious objects, so it was introduced here only for consistency with the descriptions of other theories.

In this example, the *map* transformation was used, which means that consciousness is reduced to physical processes in the spirit of reductive physicalism. If we intended to present consciousness as an entity of a different substance than $Universe$, then we would have used $transitionMind$ instead of *map* in (44a).

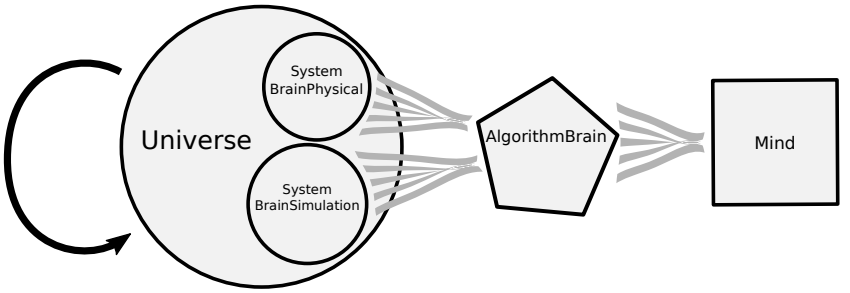

Figure 10: Illustration of functionalism.

In some panpsychic approaches, it may be possible that certain physical processes do not produce any subjective experience – in such a case we assume that *map* returns an empty state (no facts). The Integrated Information Theory framework mentioned earlier can be considered one of the possible implementations of the *map* transformation.

In the alternative formulation to (44a),

$$\forall_{s \subseteq Universe(t)} \; Mind_s = map(s), \tag{44b}$$

a certain subjective experience is also generated for every part of the $Universe$, but the $Mind_s$ objects are not dependent on time – they are like still frames associated with particular parts of $Universe$ in the time moment t. The use of the quantifier over all possible *limitscope* transformations in (44a) allows for $Mind$ to be parametrized with time – so that $Mind$ is not a set of unrelated states, as in (44b).

While most theories take a stance regarding substances and the ontological status of consciousness, panpsychism abstracts from such considerations. The number of different substances will depend on whether a *map* or a *transition* transformation is used in (44a) and (44b). If *map* is used, then there will be one substance, and if *transition* is used, then there will be two substances.

3.5. Functionalism

According to *functionalism*, a conscious mind arises from the execution of a certain algorithm, irrespectively from the underlying implementation of that algorithm; this is called *multiple realizability* [15]. Functionalism falls close to panpsychism in that many processes can be said to execute algorithms, and thus a certain kind of mind might be theoretically ascribed to almost everything in the universe. It is also similar to epiphenomenalism in that the created mind does not influence the workings of the world, with the main difference being no claim about the distinction of substances in functionalism.

In mappism, functionalism can be described as follows (see also Fig. 10):

$$Universe(t+1) = transition(Universe(t)) \tag{45}$$
$$System_x(t) = limitscope_x(Universe(t)) \tag{46}$$
$$Algorithm_x(t) = mapAlgorithm_x(System_x(t)) \tag{47}$$
$$Mind_x(t) = mapMind(Algorithm_x(t)) \tag{48}$$
$$\forall_{y,z}\ Algorithm_y(t) = Algorithm_z(t) \implies Mind_y(t) = Mind_z(t) \tag{49}$$

We assume a set of entities x, each of which possesses a mind. Each such entity is associated with its respective $System_x$ object, which is defined as a certain *limitscope* of the $Universe$ (46). $System_x$ can represent any part of the universe, including, e.g., a brain or a computer simulation of a brain. $System_x$ is then mapped to $Algorithm_x$ (47), which denotes a high-level algorithmic procedure realized by that fragment of the $Universe$. $Mind_x$ is a result of mapping this algorithmic procedure to some mind using the $mapMind$ transformation (48). If any two systems in the $Universe$ are mapped to the same $Algorithm$, then – assuming that $mapMind$ is deterministic – they will, consistently with multiple realizability, produce exactly the same $Mind$ (49). This last equation is the consequence of equations (47) and (48), and is only presented to emphasize that the principle of multiple realizability is reflected in the mappism description.

Functionalism and certain variants of panpsychism, such as Information Integration Theory (IIT) [36], are very similar but differ in one subtle detail, which we find important to elaborate upon. In IIT, a certain abstract description of a system is constructed. Thus, the general scheme of IIT would look in mappism like this:

$$System_x(t) = limitscope_x(Universe(t)) \tag{50}$$
$$Algorithm_x(t) = mapAlgorithm(System_x(t)) \tag{51}$$
$$Mind_x(t) = mapMind(Algorithm_x(t)). \tag{52}$$

The important difference is the lack of the index in the $mapAlgorithm$ transformation, which means that there exists only a single mapping, and it is used for every physical object. In contrast, in functionalism, many algorithmic interpretations of a physical system are considered ($mapAlgorithm_x$ in (47)). This is because different physical systems that realize the same function (e.g., a computer simulation of the brain and the real brain) could not be mapped by the same transformation from the low-level physical facts to the same algorithm (unless such a single transformation is designed and fine-tuned to handle both cases specifically). It is by virtue of considering all possible algorithmic interpretations of systems that functionalism has the property of multiple realizability.

Functionalism, similarly to panpsychism, abstracts from the ontological considerations about substances. The number of different substances will depend on whether a *map* or a *transition* transformation is used in (48) – if $mapMind$ stays as it is, then there will be one substance, but if some $transitionMind$ were used instead, then there would be two substances. Another *map* transformation (i.e., $mapAlgorithm$) is used in (47), but since the algorithmic description $Algorithm_x$ of some system is

only a higher level of description of this system rather than some independent entity, changing this $mapAlgorithm$ transformation to some $transitionAlgorithm$ would be unwarranted.

4. Brain, mind, and consciousness in Artificial Life

The previous section summarized classical philosophical theories of mind that attempted to describe and explain the nature of consciousness and its place in the human world. Currently, a more and more popular trend in science – due to progress in technology – is synthesizing working complex systems that in their sophistication approach the level of natural, biological systems.

The field of study that concerns synthesizing complex systems in order to study phenomena related to life is Artificial Life [2]. Working models of life-like forms are constructed in three kinds of media: software, hardware, and wetware. The field of Artificial Life is very broad and the range of phenomena that are related to life is much larger than those concerning brains and minds. One could argue that consciousness does not necessarily require life, or that it is not necessarily a property of living forms – but such arguments are a matter of definitions and we will not focus on this issue here. Instead, let us investigate whether a particular medium involved in Artificial Life research has any influence on the relationships between brains, minds, and consciousness in artificially synthesized life-like systems.

The medium that is the closest to natural (biological) life is exploited by research in wetware. Such research involves synthesizing biological cells or, in the future, organisms built (developed) from biological components. Contrary to wetware, the hardware medium seems more controllable – agents are constructed from electric and electronic devices, and potentially include sensors and actuators. Finally, the software medium appears to be the most transparent and open to inquiry: random noise is minimized, and experiments can be easily stopped, recorded, repeated, and investigated.

In this order of media involved in artificial life: wetware, hardware, software, the similarity to the biological world decreases, the cost of development decreases, potential dangers decrease (as much as humans are separate from the software world), and the flexibility of building models increases, as well as the ease of experimentation and research. What is, however, common in all these media is that life-like systems are constructed (or *synthesized*) bottom-up in order to study and create interesting phenomena, whereas traditionally, when studying biological systems, they are *analyzed* top-down in order to understand the way they work. This synthetic approach opens up new ways of research and experimentation, and it could greatly influence the research on consciousness, but we would first need a useful definition of consciousness (or mind) in order to come up with ways to measure (detect) this phenomenon. This may be as difficult as building detectors of intelligence and detectors of life, not only because of technical problems, but because of the lack of precise definitions of these phenomena.

Are there any differences between wetware, hardware, and software when it comes

to the potential to create minds or consciousnesses? The difference between hardware and software is quite fuzzy, but does wetware have any special properties that would facilitate the emergence of consciousness in addition (or contrary) to what hardware and software can offer? Since we are synthesizing life forms, we are in the engineering domain, and this usually means that creators "know what they are doing" and it is hard to expect unknown phenomena to appear when one is using perfectly known building blocks. In wetware, however, these blocks are not entirely understood, and even in software we are capable of developing systems so complex that they by far exceed human comprehension – not to mention software and hardware systems that develop and modify themselves, sometimes randomly. These considerations demonstrate that human engineers themselves may be unable to understand their own creations [20], and there is room for the unexpected emergence of phenomena like consciousness – possibly, a gradual emergence, so we could speak about a "simple mind" with a "low consciousness level", for example.

Even if we allow for such a possibility, the engineering, scientific, and technological biases make it hard to ascribe emerging complex properties to some forces outside of the scope that is inherent to the building blocks used for development. The bottom-up synthetic approach suggests that emergent phenomena are all grounded in the basic level where building blocks are used. This is in line with current scientific perspective that even complex, high-level phenomena such as intelligence or life can be somehow defined and measured without referring to mysterious, external influences.

Following this line of thought, not only feelings, emotions, and complex phenomena (like for example intelligence, free will, and consciousness) can be described and measured, but it may be possible to record them, clone (reproduce), and possibly transfer to different media. The difficulty lies in the fact that some of such notions may be subjective, and most of them operate on a very high level of abstraction far from the level of building blocks, so they can be considered an extremely lossy compression of the underlying low-level state of the system. In the following subsections, we will demonstrate how mappism can be used to address this Artificial Life software perspective where multiple levels of abstraction are involved, yet the basic level of building blocks is perfectly known, well defined and understood.

4.1. A cellular automaton universe

Let us start with a very simple example – a cellular automaton (CA): a discrete system that consists of a grid of cells [26], each cell with some state, and a set of rules which determine the state of each cell in the next step based on the current state and the states of the neighboring cells (53). Note that in a CA everything is precisely defined as mathematical functions or algorithms, so on this lowest level we have perfect and complete information about how the system evolves. "Perfect and complete" does not necessarily mean deterministic, as the rules and the way they are applied can be probabilistic, but we still assume that we know their precise definition.

Some configurations of a CA are more interesting than others; one could imagine a stagnant CA where all cells, when rules are applied to them, do not change their states.

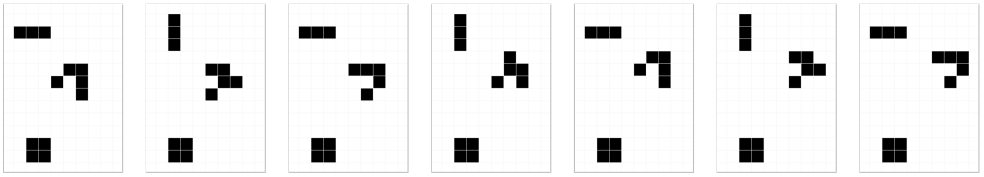

Figure 11: From left to right, seven consecutive states of a simple cellular automaton. Each cell is square and can only have one of the two states (shown as white and black). There are three black structures visible: one static, one oscillating, and one moving in the up-right direction.

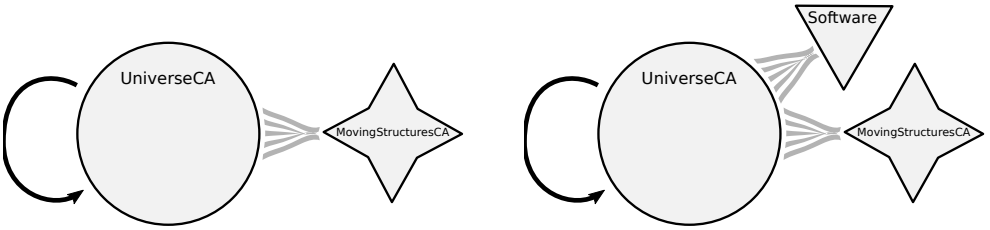

Figure 12: Visual notation of a cellular automaton universe. Left: equations (53) and (54), right: including also mapping (55).

Another possibility is a CA where all states change all the time in a seemingly chaotic way. Or perhaps there are some structures of states that move in some direction – a good illustration is the famous "game of life" CA [26] with "gliders" as an example of moving sets of neighboring cell states (Fig. 11).

Following this example, let us assume that for the purposes of some analysis, one is only interested in moving structures, therefore to describe the system we need to only describe precisely these moving structures. To reflect this shift in perspective, we change the language in which we describe the spatiotemporal dynamics of the system – instead of a low-level specification of the states of all cells in each step, we use terms like "two gliders moving in the up-left direction" or "structure S_{53} moving downwards". The new, higher level language (54) does not precisely describe the state of the system, but it may be more appropriate for the analysis because it describes emergent phenomena and ignores phenomena that are not relevant for the analysis (e.g., stationary structures). Note that in a CA, in each step the state of a cell is only influenced by the neighboring cells (the property of locality), so there is a limited propagation time of information, similar to the speed of light in our universe. We could precisely define "spatial past", "spatial future", and the concept of causality in a CA, analogously to what results from the "light cone" in our physical universe.

$$UniverseCA(t+1) = transition(UniverseCA(t)) \tag{53}$$

$$MovingStructuresCA(t) = mapCAtoMS(UniverseCA(t)) \tag{54}$$

The two languages, low- and high-level, that were introduced earlier can be illus-

trated in mappism as shown on the left in Fig. 12. Some CAs are Turing complete (computationally universal, i.e., allowing to perform any computation), so they can simulate a Turing machine [25, 32] and, in consequence, any calculation or computer program. This can be achieved in many ways – simulating electronic circuits in a CA, or simulating data structures and rules that process these data structures. The fact that a CA can be used to run software bridges the gap between a seemingly trivial automaton and the world of algorithms and high level software concepts, so we can add (55) and extend our graph as shown on the right in Fig. 12. The mapping between the CA and software can be performed without any loss of information and unambiguously in both ways if this is the intent, but this would not be very practical. There are many CAs that perform the same specific function of a given software/algorithm, and we are primarily interested in the logic and behavior of the algorithm, and not the specific layout (e.g., specific location and rotation of cell structures) of the underlying CA. Therefore, similarly to the "moving structures" language, the algorithm-level description of the behavior of a CA is higher level and may lose some irrelevant information about how specific cells in a CA behave.

$$Software(t) = mapCAtoSoft(UniverseCA(t)) \tag{55}$$

Note that in this simple description, we never did refer to the medium that actually runs the CA. We might have assumed that the CA is simulated on some computer (as it usually is – using some software, obviously operating on a different, lower level than the software defined in (55)), but the underlying substrate could as well be mechanical, chemical, or electronic (without the need for software). The underlying medium was not included in this concise three-equation description.

In Sect. 3 we briefly characterized major classical theories of mind, but we had not precisely defined high-level objects that were used (e.g., $Brain$, $Mind$, $Universe$) and various complex mappings that were used to transform these objects. We left the precise definitions of these objects and transformations to physics, chemistry, biology, psychology, and other fields involved in this discussion. In this section, however, since we introduced a cellular automaton universe, we will at least provide a few additional pieces of information on objects and transformations that were mentioned in (53), (54), and (55), just to give an example of a more explicit description and to make it clear that these functions can be formally defined down to the lowest level if needed.

We have three languages involved in this system, used to describe the low-level CA (let us denote it as $Cells_{Language}$), "moving structures" ($MovingStructures_{Language}$), and software ($Software_{Language}$). Following the definitions in Sect. 2.1, an example of a fact in $Cells_{Language}$ would be "The cell number 123 has the state number 4", assuming all cells and all cell states have unique numbers (IDs). An example of a fact in $MovingStructures_{Language}$ would be "structure number 678 is moving in direction number 9", and this would obviously require defining all shapes of structures that are of interest and uniquely identifying their possible directions of movement. Finally, an example of a fact in $Software_{Language}$ depends on how this language is defined and what the purpose of this language is – a fact could be an (input, output) pair (i.e., what output or behavior of the CA results from a particular input), but it could also

be some partial information about the logic of the algorithm that the CA executes.

We can define the three languages more formally as

$$\begin{aligned} Cells_{Language} &= \{(c,s) \ : \ c \in \{1,..,C\}, s \in \{1,..,S\}\} \\ MovingStructures_{Language} &= \{(m,d) \ : \ m \in \{1,..,M\}, d \in \{1,..,D\}\} \\ Software_{Language} &= \{(i,o) \ : \ i \in \{1,..,I\}, o \in \{1,..,O\}\} \end{aligned}$$

where C is the total number of cells in the CA, S is the number of unique states, M is the number of moving structures of interest, D is the number of directions they can move in (assuming for simplicity that there is some common set of directions and we do not have to separately define directions and movement stages for each structure type), I is the number of unique software inputs, and O is the number of unique outputs. Note that analogously to structure types and directions, input-output pairs are higher-level concepts and they do not belong to $Cells_{Language}$ – instead, they are facts in the software realm, and sets of such facts constitute software states that are the appropriately transformed CA states (55).

To summarize, each of the three languages is defined as a set of all distinct facts that can be expressed in its representation, so we can write respectively

$$\begin{aligned} Cells_{Fact} &\in Cells_{Language} \\ MovingStructures_{Fact} &\in MovingStructures_{Language} \\ Software_{Fact} &\in Software_{Language} \end{aligned}$$

where X_{Fact} is any fact from the language $X_{Language}$.

States that are expressible in each language are sets of facts, i.e., subsets of a given language:

$$\begin{aligned} Cells_{State} &\subset Cells_{Language} \\ MovingStructures_{State} &\subset MovingStructures_{Language} \\ Software_{State} &\subset Software_{Language} \end{aligned}$$

where X_{State} is any state expressed in the language $X_{Language}$.

Finally, let us characterize objects and transformations used in (53), (54), and (55) in terms of their domains and codomains:

$$\begin{aligned} UniverseCA &: \mathbb{N} \to \mathcal{P}(Cells_{Language}) \\ MovingStructuresCA &: \mathbb{N} \to \mathcal{P}(MovingStructures_{Language}) \\ Software &: \mathbb{N} \to \mathcal{P}(Software_{Language}) \\ transition &: \mathcal{P}(Cells_{Language}) \to \mathcal{P}(Cells_{Language}) \\ mapCAtoMS &: \mathcal{P}(Cells_{Language}) \to \mathcal{P}(MovingStructures_{Language}) \\ mapCAtoSoft &: \mathcal{P}(Cells_{Language}) \to \mathcal{P}(Software_{Language}) \end{aligned}$$

where we assume that time t is discrete (hence the set of natural numbers $\mathbb{N}$), and use $\mathcal{P}(X)$ to denote the power set of set X, i.e., the set of all subsets of X.

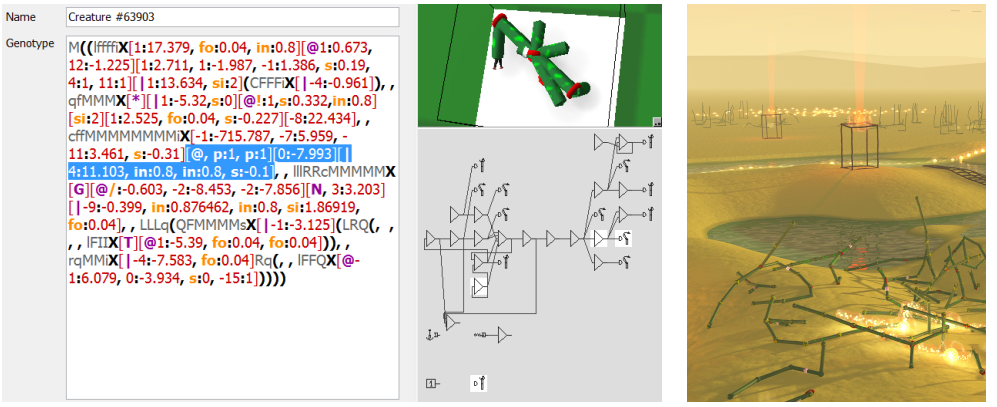

Figure 13: A sample simulation of creatures, Framsticks [21, 22]. Left: a single creature with its "genotype" which encodes body, brain, receptors, and effectors, and which can be modified by appropriate genetic operators. Right: a view of the virtual world with multiple creatures interacting.

4.2. A coarse simulation of biological creatures

Let us now consider another kind of an artificial life simulation – a coarse simulation model of biological organisms. Contrary to cellular automata described above, which were highly homogeneous and simple to define, in this simulation model biological creatures are simulated on a standard computer (a CPU with memory) or on some Turing machine equivalent using high-level building blocks: simulated neurons and various kinds of body parts. Such simulated creatures (agents) are also equipped with simulated sensors (receptors) and actuators (muscles), so that they can process information acquired by sensing the virtual environment they are simulated in and perform actions using simulated muscles. Some receptors and effectors can be used for communication between agents, and indirect communication by influencing the virtual environment (stigmergy) is possible as well. Physical forces like gravity and friction are also simulated in this virtual world. We may also assume that these agents can change – they can adapt to their environment and potentially evolve (example shown in Fig. 13).

Since each simulated agent has a neural network consisting of neurons that use some more or less accurate model of biological neurons, we will call this network a (simulated) brain, and the mechanical part of a creature – a (simulated) body. Compared to our world which seems to be governed by a small set of physical rules (similarly to a cellular automaton), such a simulation is *coarse* and heterogeneous – nearly every process is handled by software as a special case and a special procedure, because this is a high-level model: neurons are different from body parts, brain (an artificial neural network) and body (a structure made from small finite elements) are simulated in a completely different way, reproduction is a specific procedure, energetic balance of creatures is handled as a separate procedure with its own logic, etc.

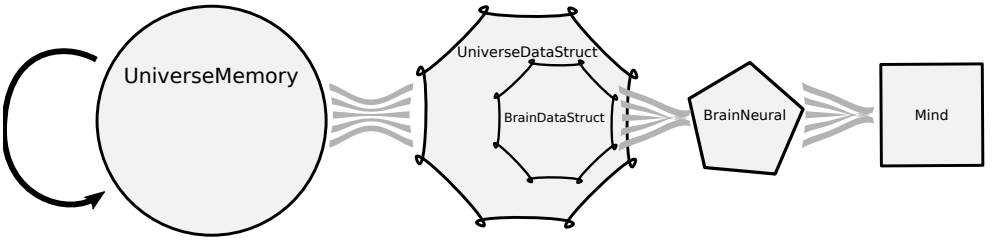

Figure 14: Visual notation for simulated biological universe.

Despite the fact that the model is coarse and heterogeneous, it is a working computer program and so it is precisely described in terms of algorithms and mathematical functions. Let us consider a memory of a computer as our basic level of analysis. Then we can describe this system and its individual agents x using mappism in the following way, illustrated in Fig. 14:

$$UniverseMemory(t+1) = transition(UniverseMemory(t)) \tag{56}$$
$$UniverseDataStruct(t) = remap(UniverseMemory(t)) \tag{57}$$
$$BrainDataStruct_x(t) = limitscope_x(UniverseDataStruct(t)) \tag{58}$$
$$BrainNeural_x(t) = mapBDStoBN(BrainDataStruct_x(t)) \tag{59}$$
$$Mind_x(t) = mapBNtoMind(BrainNeural_x(t)) \tag{60}$$

Just as in the previous section where we defined objects and functions for a cellular automaton, let us have a closer look at these functions, starting with languages used on different levels of description.

$$Memory_{Language} = \{(m, b) \ : \ m \in \{1, .., M\}, b \in \{0, 1\}\}$$
$$DataStruct_{Language} = \{(i, v) \ : \ i \in \{1, .., I\}, v \in Integers32bit \cup Floats64bits \cup \ldots\}$$
$$Neural_{Language} = \{(n, a) \ : \ n \in \{1, .., N\}, a \in Real\} \cup T$$
$$Mind_{Language} = \{\text{SU\_Hungry}, \text{SU\_Happy}, \text{SU\_FoodNearby}, \ldots\}$$

Since we have chosen computer memory to be our base level of description, we are not concerned with lower levels (i.e., the underlying electronics and physics, or some other substrate on which the computing machinery works). The $Memory_{Language}$ describes values of bits in memory that can store M bits. The $DataStruct_{Language}$ is the level of description most computer programmers deal with – the level of variables organized in data structures (we assume I variables) that store integer or floating-point 32-bit or 64-bit numbers, Boolean values, arrays, strings of characters, matrices, memory addresses, etc. The $Neural_{Language}$ describes outputs of simulated neurons, assuming that an agent has N neurons, and that activations a of neurons can be represented in some available implementation of real values. The T set describes the topology of the neural network and any additional information needed to understand how the network works, such as different types of neurons or receptors, etc.

Finally, $Mind_{Language}$ is a set of interpretations of activation patterns in the simulated universe (hence the SU prefix) – even though names such as SU_Hungry and

SU_Happy might suggest that they come from our (human) world, they are in fact terms that could be assigned to specific neural activations or the corresponding behaviors by (possibly intelligent and communicating) creatures that "live" in that virtual environment. There is obviously no direct matching between SU_Hungry and our (human) hunger, and there is also no direct matching between SU_Hungry and the behavior that a human researcher studying behaviors of these simulated agents could assign to their neural activations that seem to make these agents engage in food foraging behavior. Again, SU_Hungry is a name for a specific state of a brain of a simulated agent; this may be subjective to each agent, but if these agents develop communication skills and some kind of shared signs (not to mention a complex language), in some scenarios it would be highly beneficial to share these names between agents. As these agents are themselves similar and experience similar conditions, they also share similar states, so communicating such states using a consistent set of signs would be in many cases advantageous. Assuming that such self-awareness, the ability to distinguish oneself from others, the memory of past self, and the ability to predict (model) the consequences of one's actions are sufficient for (simulated) consciousness (cf. [20]), we could also call such states (simulated) qualia. Note that in the previous sentence, "simulated" does not mean that we are imitating or modeling human consciousness or qualia. It means that these phenomena would originate from (or exist in) $UniverseMemory$, and we traditionally call processes that run there "simulations". An alternative adjective to "simulated" could be "artificial".

Note that in (60), the mapping function $mapBNtoMind$ is independent from x, i.e., there is one such function for each way of interpreting (defining) qualia. This function may be very difficult to define, perhaps even impossible to be defined by some agent (a human, a simulated creature, or some external or internal observer); the mappism description merely indicates that such a function exists and can be constructed or discovered. The existence of such a function does not contradict the fact that qualia are subjective, it only assumes that they can be defined and specified. If we wanted the qualia to be subjective to the degree of being impossible to compare between agents, we could introduce separate languages for each agent (i.e., $Mind_{x\_Language}$) or separate facts describing qualia specific to each agent. Humans, however, share the terms used to describe subjective qualia because of communication, similarity of organisms within species, and "low resolution" (high compression or high loss of information) of our languages – even though "my happiness is not the same as your happiness" and "my happiness now is not the same as ten minutes ago", the word "happiness" still exists and can be applied to anyone.

The argument of the $mapBNtoMind$ function, i.e., the information about the state of a brain expressed in $Neural_{Language}$ should be complete – it should represent all the influences that affect (even unconsciously) the brain and have a chance to influence mind. In case of humans, this would mean that we might need to also consider states of the enteric nervous system, parts of the neural system that are influenced by various bacteria such as the gut microbiome, etc.

The $Mind_{Language}$ set was defined here as consisting of individual qualia that originate from $UniverseMemory$, so a state of an agent's mind would be a subset of $Mind_{Language}$. This is the most trivial and rudimentary representation of

the mind state, and there exist many more general models. One example would be associating the degree of intensity, or truth, with each quale, so instead of $\{\text{SU\_Hungry}, \text{SU\_Happy}, \text{SU\_FoodNearby}, \ldots\}$ we would model the state of the mind as $\{(\text{SU\_Hungry}, 0.9), (\text{SU\_Happy}, 0.2), (\text{SU\_FoodNearby}, 0.1), \ldots\}$. More realistic modeling of qualia might require using more complex models with additional (possibly mutually constrained) attributes; such models are available and are employed in Artificial Intelligence whenever complex representations of knowledge are needed.

Facts and states can be defined for each of the four introduced languages analogously to the definitions in the previous section: facts are individual elements from the language sets, and states are subsets of these sets. Similarly, domains and codomains for objects and transformations in equations (56)–(60) can be deduced directly from their arguments and results, still we list them explicitly below:

$$
\begin{aligned}
UniverseMemory &: \mathbb{N} \to \mathcal{P}(Memory_{Language}) \\
UniverseDataStruct &: \mathbb{N} \to \mathcal{P}(DataStruct_{Language}) \\
BrainDataStruct_x &: \mathbb{N} \to \mathcal{P}(DataStruct_{Language}) \\
BrainNeural_x &: \mathbb{N} \to \mathcal{P}(Neural_{Language}) \\
Mind_x &: \mathbb{N} \to \mathcal{P}(Mind_{Language}) \\
transition &: \mathcal{P}(Memory_{Language}) \to \mathcal{P}(Memory_{Language}) \\
remap &: \mathcal{P}(Memory_{Language}) \to \mathcal{P}(DataStruct_{Language}) \\
limitscope &: \mathcal{P}(DataStruct_{Language}) \to \mathcal{P}(DataStruct_{Language}) \\
mapBDStoBN &: \mathcal{P}(DataStruct_{Language}) \to \mathcal{P}(Neural_{Language}) \\
mapBNtoMind &: \mathcal{P}(Neural_{Language}) \to \mathcal{P}(Mind_{Language})
\end{aligned}
$$

As in the previous section, we assume that time t is discrete, and $\mathcal{P}(X)$ is the power set of set X.

4.3. Recurrent multi-layer artificial neural networks

Another area where neural networks are often simulated is artificial intelligence and, in particular, machine learning. The most classical architecture of artificial neural networks used in supervised learning is a multi-layer feed-forward architecture, where data are fed into the input (first) layer, and then the signals get propagated through layers until they reach the output (last) layer. Adjacent layers are connected fully, but there are no recurrent (i.e., going back from neuron outputs to previously processed inputs) connections in this architecture. In a typical usage scenario, layers are used to transform representations of data, going from the original space of attributes that describe data instances (input layer) to the desired decision attributes (output layer). The dimensionality of the output layer is usually much lower than the dimensionality of the input. A similar goal is achieved by autoencoders that discover (learn) a smaller representation (a simpler or higher-level language) of original data while trying to preserve the overall characteristics of the data as much as possible.

Let us consider a specific example that will be related to our discussion of qualia from the previous section 4.2 on simulated biological creatures. The data set will come from a survey of a thousand people asking for their income, age, martial status, family, work, health, etc., and finally for their satisfaction in life (with only two choices: happy or unhappy). The artificial neural network in the supervised learning scenario could be used to learn, for the i-th example in the dataset, to predict the decision attribute (happy or unhappy – using one or two neurons in the output layer) based on the conditional attributes (income, age, etc.). The input layer would have many more neurons than the output layer – this kind of architecture encourages the network to come up with a more and more compressed representation of the target concept (i.e, happy or unhappy) in each subsequent layer. This should remind us of the *map* function and indeed, this is the correct analogy. The ability of subsequent neural layers (62)-(65) to transform representations and reach higher and higher levels of abstraction is in fact one of the major reasons for the success of deep learning – think about processing images and going from information about the colors of individual pixels, through edges, then curves, then shapes, symbols, letters, words, sentences, paragraphs, and ultimately to meanings, emotions, feelings, and aesthetic impressions.

$$Example(i) = limitscope_i(UniverseDataSet) \tag{61}$$

$$Layer1Outputs(i) = map1(Example(i)) \tag{62}$$

$$Layer2Outputs(i) = map2(Layer1Outputs(i)) \tag{63}$$

$$Layer3Outputs(i) = map3(Layer2Outputs(i)) \tag{64}$$

$$Layer4Outputs(i) = map4(Layer3Outputs(i)) \tag{65}$$

Would the happy/unhappy neuron output have anything to do with consciousness? Since that output (the last layer of neurons (65)) seems to use the language of qualia, could we call this last layer a $Mind$? Note that the notion of happiness here differs in two important aspects from the notion of happiness (SU_Happy) discussed in the previous section. First, the goal of supervised learning as presented here is to make the artificial neural network learn the concept described by the data set (in this case, the relationship between the values of the conditional attributes and happiness). So the goal of the artificial neural network is precisely to capture the notion of human happiness as the survey data define it, not some other (simulated) happiness of some (simulated) agents. The language that the output layer uses is precisely the language of human (un)happiness, so if anything, these are human qualia. Second, it does not seem reasonable to speak about any form of genuine consciousness of the neural network itself because the network is feed-forward, i.e., it has no feedback loops and is unaware of its own state [29]; the state only depends on the dataset, and in particular, on a single example that was fed into the network.

If we, however, consider another artificial neural network architecture, e.g., a deep recurrent architecture with many layers and multiple recurrent connections, then such networks are capable of changing their behavior in time based on their own current and past internal states. Embedding such a network in some environment (so that it is embodied and situated), making it able to adapt, interact with the environment, observe its own actions, and perhaps communicate with other networks would make

questions whether it possesses consciousness and qualia less obvious to answer. Similarly to the configuration described in the previous section 4.2, representations of states of such a network would then concern its virtual environment and not our human world (dataset). Therefore, the answer to such a question would depend on how the notions of consciousness and qualia are precisely defined.

Interpreting the behavior of such a network from the perspective of functionalism, one would say that the network may be conscious in a similar sense as humans are if it realized a certain procedure that is complicated enough [30]. On the other hand, from the point of view of Information Integration Theory [36, 37], which is a panpsychic approach, such a network could be conscious if implemented in wetware or in appropriate hardware, but not if simulated as software.

The perspective and experiences of an artificial intelligence and artificial life researcher encourages to define and study consciousness in terms of specific abilities or properties of the information-processing system. Such properties include memory, the distinction between self and everything else, the model of self and the successful prediction of the results of actions performed by self, and the ability to sense self through feedback loops and to make this information influence future states of self. These properties and abilities are easily attainable by contemporary software and hardware systems. Moreover, artificial life allows to build working models that implement any classical theory of consciousness by appropriately designing objects and transformation functions, and then perform simulation (perhaps evolutionary) experiments to see how well their results agree with our reality. For example, simulated agents with shared memory, shared partial brain, shared mind, or two artificial neural networks for each agent ("soul" and "brain") that cannot communicate directly or can only communicate one way, are possible to implement easily. Additionally, simulated worlds and the real world can be bridged [20] by allowing for some range of interactions (e.g., connecting $BrainNeural_x$ (59) or $Mind_x$ (60) to our world by interfaces such as a keyboard, a camera, a microphone, a robot arm, or a screen). This would enrich the information in feedback loops and increase their number; otherwise such feedback loops would be limited to agents themselves and interactions between agents in the simulated world. However, since consciousness is far from being fully understood, such experiments require careful ethical considerations and the analysis of consequences first.

4.4. Universes embedded one in another

In Sects. 4.1, 4.2, and 4.3 we discussed artificial life software models, so we had full access to encoded information and full knowledge about every fact in every considered language. Still, the *map* transformations may make such considerations difficult because, especially towards higher level of abstraction, there is a considerable loss of information, and these mappings can be defined in many different ways (recall the notion of SU_Happy). This is what makes the discussion about mind and consciousness so complicated – especially in real life, where very complex mappings are required, high-level notions are used, and it is difficult to define them precisely, unambiguously,

and even more so, objectively.

Reflecting on Sect. 4.1, one can see that a cellular automaton (CA) can perform equivalently to a Turing machine [25, 32], and that a CA itself can be (and often is) simulated in software on a traditional computer:

$$UniverseCA(t) = mapSofttoCA(Software(t))$$

Since a CA can be simulated in software and a CA itself can simulate running software, it means that a CA can simulate, through a software layer, another CA, and this chain can be made arbitrarily long by repeatedly using the mapping functions.

Recall that in Sect. 4.2 we started our description of the biological simulation from the base level of software, so the two mappism descriptions from Sects. 4.1 and 4.2 can be merged – we can bridge (55) and (56) in the following way:

$$UniverseMemory(t) = remap(Software(t))$$

where $Software(t)$ was the highest-level language in the "CA universe", but the lowest-level language in the "biological simulation universe". When languages are compatible, such mappings can be merged and composed – this corresponds to substitution in mathematics while ensuring that types of arguments of the relevant functions are correct (as explained in Sect. 2.7).

The concept of universes embedded one in another has been discussed in more detail in [20]. One conclusion we may draw from the above considerations is that once we have some object and its transition function, we can forget about lower levels and mappings. This is because each object constitutes some realm – an example could again be the cellular automaton universe: it does not matter from the subjective point of view of the behavior of the CA what the underlying substrate is and how many mappings were used to reach the CA level.

Let us finish this section with a few observations regarding imagination, consciousness, and qualia. In Sect. 4.2 we described the universe of simulated creatures, and in (60) we introduced a function that defines high-level concepts like "hunger" or "happiness", assuming that these simulated creatures are able to sense and identify their self (i.e., their own states and the influence of their actions on the world) and are sophisticated enough to build a model of self (predict consequences of their actions, remember past experiences, etc.). Such creatures, for whatever reason, might be able to come up with a concept of a cellular automaton (either because they observed a CA as a part of their simulated universe, or as an abstract concept). They would be then able to imagine specific states of a CA and simulate a CA in operation – i.e., they could compute the *transition* function of the imaginary CA as shown at the bottom of Fig. 15. Their CA simulation would likely proceed in a different time scale, because complex mappings may deteriorate performance – our civilization tries to avoid this cost by bringing computation closer and closer to the low-level physical realm [20]. The language that the simulated creatures would use to describe the imaginary CA would be equivalent to, or identical with, the language that is generally used to describe a CA ($Cells_{Language}$ in earlier equations, even though the full definition of a CA is slightly more complex than our examples here).

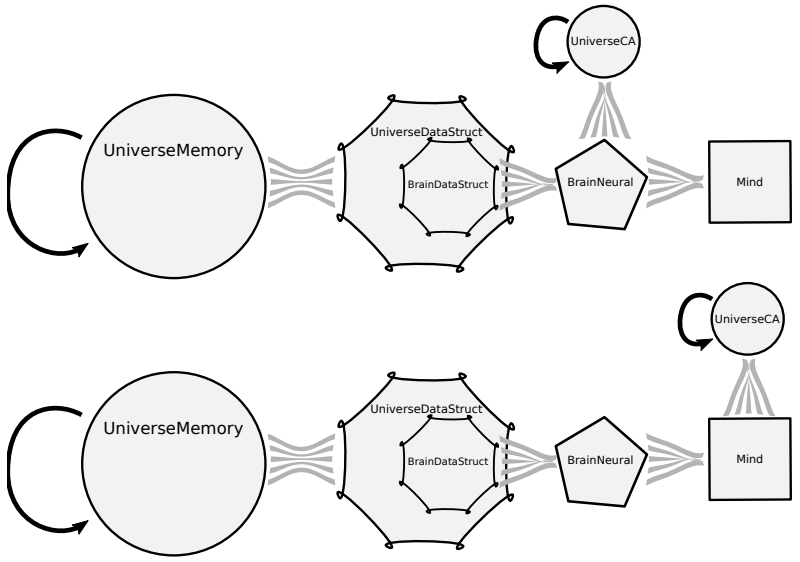

Figure 15: Illustration of simulated biological creatures simulating a working cellular automaton (CA). Top: states of a CA directly reflecting (i.e., mapped from) states of a brain. Bottom: conscious imagination of a working CA.

Assuming that the biological simulation runs on some CA as the underlying substrate, the situation described above makes it obvious that there is a difference between the CA imagined by simulated creatures in their artificial neural networks, and the CA that underlies the simulation of these creatures, even though the language and the transition function of both CAs could be identical. One could say that the underlying CA *exists*, while the CA imagined by simulated creatures is *just an imagination* of the concept, even if the creatures are actually imagining specific states of the CA and actually calculating values of its *transition* function. We could be tempted to say that both CAs are different *instances* of the CA concept, but is the CA imagined by these creatures (along with its precisely calculated states) really an *instance*?

Answering this question is important, because a very similar example would be even more intriguing: the simulated creatures imagining another universe (with some other creatures), specifying the state of such an universe, and calculating its *transition* function along with other functions that led to high-level notions like SU_Happy. Such a configuration reminds us of fundamental questions concerning our reality, discussions whether this reality is a simulation, whether simulated creatures can be conscious, whether simulated, intelligent, and conscious creatures can discover that they are a simulation, etc.

This is where we should restrain from abusing our human language and restrain from using the same, high-level words like “consciousness”, “mind”, or “simulation” for different concepts, as it causes confusion and makes problems seem more difficult. The difference between the “imagined CA” and the “underlying CA” in the example

given above is described precisely by the relevant mappism functions, and discussing whether these CAs should be called *instances* is fruitless unless one first precisely defines the word "instance", which would result in a number of subjective statements referring to other vague, high-level terms like "real", "object", or "existence" that would again require precise and objective definitions.

Where we have multiple levels of embedding of universes with multiple mappings and languages (representations), it is difficult to use our human language precisely (see also the discussion in Sect. 3.3 of [20]). Instead of trying to describe complex phenomena like mind and consciousness using high-level human-language words that have many meanings and suffer from a significant loss of information, we recommend using simple functions provided by mappism to describe relationships between well-defined entities that are involved in the emergence of these phenomena. Artificial life models may be of great help here due to their explicit nature, the ease of manipulation, and the ability to be formalized and precisely described. Such simulated models, diverse environments, and evolutionary pressures may lead to the emergence of multiple abstraction layers and complex phenomena such as communication, language, and consciousness. Working models will not only facilitate the study of these phenomena and related concepts like subjective perception, qualia, and philosophical zombies, but also allow researchers to introduce and use precise definitions when describing these concepts.

5. Summary

The problem of mind and consciousness is one of the most fundamental problems considered in philosophy. Many theories on this subject were presented in the past, but because of their complexity and descriptions in imperfect human languages, in order to avoid ambiguities, these theories often require verbose definitions. Such extensive definitions and multiple variants make the theories hard to compare and hard to unambiguously comprehend. This is because any non-formal language leaves a lot of space for inaccuracies and interpretation, and in philosophy, many informally used words (e.g., soul, mind, consciousness, reality, truth, purpose) may have multiple, alternative definitions.

In this paper, we have presented a way to talk about and compare theories of mind in the unified framework of *mappism*. Mappism embraces the perspective in which the universe (and everything in it) can be represented in terms of mathematical concepts – variables, sets, functions, quantifiers, etc. We have proposed a set of tools – including transformations, which are specialized types of functions – which allow for a precise description of relations between different objects such as $Universe$, $Brain$, or $Mind$. Major classical, philosophical theories of mind have been presented and their formulations in mappism have been proposed in Sect. 3. We have also shown how mappism can be used to describe consciousness that could potentially emerge in simulation in artificial life experiments (Sect. 4), where mappism objects and transformations can be defined more concretely and in full detail.

Mappism in itself is not a theory of mind and consciousness; instead it is a way

to describe relationships within a universe, and in particular, relationships between brains, minds, and realities. As mappism uses formal means, it can be helpful not only in identifying similarities between theories and their variants as demonstrated in Sect. 3 of this paper, but also in exploring their consequences, or revealing inconsistencies and contradictions. Employing formal means to describe theories forces one to be specific, and allows one to manipulate mathematical symbols and functions according to well-defined mathematical operations, such as "solve for x", "substitute x with y", subtracting a set, decomposing a composite function, etc. This also allows mappism to unambiguously describe popular concepts like China brain, Chinese room, or philosophical zombies, and to bridge theories that belong to different fields of science – from physics, through chemistry and biology, to computer science, neuroscience, and psychology.

Another application area of mappism is education and – more generally – the popularization of knowledge, where it can be used to present theories and differences between them both in a formal and a visual way. Mappism could also be a good choice whenever there is a need to represent various theories in a way that is easy for computers to process. Such a marriage of philosophy and computer science can be found in a slowly emerging field of computational philosophy [13, 27, 31]. Finally, while mappism as it is can be used to formally describe and unify existing theories of mind and consciousness, it can be further extended if necessary.

References

[1] Adams D. *The Hitchhiker's Guide to the Galaxy.* Pan Books, 1979.

[2] Bedau M. A. Artificial life: organization, adaptation and complexity from the bottom up. *Trends in cognitive sciences*, 7(11):505–512, 2003.

[3] Buckingham G., Michelakakis E. E., and Rajendran G. The influence of prior knowledge on perception and action: Relationships to autistic traits. *Journal of Autism and Developmental Disorders*, 46(5):1716–1724, May 2016. URL: `https://doi.org/10.1007/s10803-016-2701-0`, `doi:10.1007/s10803-016-2701-0`.

[4] Chalmers D. *The Conscious Mind: In Search of a Fundamental Theory.* Oxford University Press, Inc., New York, NY, USA, 1996.

[5] Chalmers D. J. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219, 1995.

[6] Chalmers D. J. Panpsychism and panprotopsychism. In *The Amherst Lecture in Philosophy 8*, pages 1–35, 2013. URL: `http://www.amherstlecture.org/chalmers2013/`.

[7] Churchland P. M. *Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind.* MIT Press, Cambridge, MA, USA, 1st edition, 1988.

[8] Coren S. and Girgus J. S. *Seeing is deceiving: The psychology of visual illusions.* Lawrence Erlbaum, 1978.

[9] Dennett D. C. *Consciousness Explained.* Penguin Books, 1991.

[10] Dennett D. C. *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness.* MIT Press, 2005.

[11] Descartes R. *A Discourse on Method.* 1637.

[12] Eysenck M. W. and Keane M. T. *Cognitive Psychology: a student's handbook.* Psychology Press, 6th ed. edition, 2010.

[13] Fitelson B. and Zalta E. N. Steps toward a computational metaphysics. *Journal of Philosophical Logic*, 36(2):227–247, 2007.

[14] Gallagher S. *Phenomenological Approaches to Consciousness*, chapter 50, pages 711–725. Wiley-Blackwell, 2017. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119132363.ch50, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119132363.ch50, doi:10.1002/9781119132363.ch50.

[15] Jackendoff R. *Consciousness and the computational mind.* The MIT Press, 1987.

[16] Kang M.-S., Hong S. W., Blake R., and Woodman G. F. Visual working memory contaminates perception. *Psychonomic Bulletin & Review*, 18(5):860–869, Oct 2011. URL: https://doi.org/10.3758/s13423-011-0126-5, doi:10.3758/s13423-011-0126-5.

[17] Key B. Fish do not feel pain and its implications for understanding phenomenal consciousness. *Biology & Philosophy*, 30(2):149–165, Mar 2015. URL: https://doi.org/10.1007/s10539-014-9469-4, doi:10.1007/s10539-014-9469-4.

[18] Kim J. *Philosophy of Mind.* Westview Press, 1996.

[19] Koch C., Massimini M., Boly M., and Tononi G. Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, 17:307–321, 2016. doi:10.1038/nrn.2016.61.

[20] Komosinski M. Universes and simulations: civilizational development in nested embedding. *Foundations of Computing and Decision Sciences*, 43(3):181–205, 2018. doi:10.1515/fcds-2018-0010.

[21] Komosinski M. and Ulatowski S. Framsticks: Creating and understanding complexity of life. In Komosinski M. and Adamatzky A., editors, *Artificial Life Models in Software*, chapter 5, pages 107–148. Springer, London, 2nd edition, 2009. URL: http://www.springer.com/978-1-84882-284-9.

[22] Komosinski M. and Ulatowski S. Framsticks web site, 2018. URL: http://www.framsticks.com.

[23] Laureys S., Gosseries O., and Tononi G. *The neurology of consciousness: cognitive neuroscience and neuropathology*. Academic Press, 2015.

[24] Leibniz G. W. Monadology. In *Philosophical Essays*. Indianapolis, Hackett, 1989.

[25] Lindgren K. and Nordahl M. G. Universal computation in simple one-dimensional cellular automata. *Complex Systems*, 4(3):299–318, 1990.

[26] Mainzer K. and Chua L. *The Universe as Automaton: From Simplicity and Symmetry to Complexity*. SpringerBriefs in Complexity. Springer-Verlag Berlin Heidelberg, 2012. URL: https://www.springer.com/gp/book/9783642234767, doi:10.1007/978-3-642-23477-4.

[27] Mayner W. G. P., Marshall W., Albantakis L., Findlay G., Marchman R., and Tononi G. PyPhi: A toolbox for integrated information theory. *PLOS Computational Biology*, 14(7):1–21, 07 2018. URL: https://doi.org/10.1371/journal.pcbi.1006343, doi:10.1371/journal.pcbi.1006343.

[28] Nagel T. What is it like to be a bat? *Philosophical Review*, 83(October):435–50, 1974.

[29] Peters F. Consciousness as recursive, spatiotemporal self-location. *Psychological Research PRPF*, 74(4):407–421, 2010.

[30] Putnam H. Minds and machines. In Hook S., editor, *Journal of Symbolic Logic*, pages 57–80. New York University Press, 1960.

[31] Rapaport W. J. and Kibby M. W. Contextual vocabulary acquisition as computational philosophy and as philosophical computation. *Journal of Experimental & Theoretical Artificial Intelligence*, 19(1):1–17, 2007.

[32] Rendell P. Turing universality of the game of life. In *Collision-based computing*, pages 513–539. Springer, 2002.

[33] Schilling M. and Cruse H. The evolution of cognition – from first order to second order embodiment. In Wachsmuth I. and Knoblich G., editors, *Modeling Communication with Robots and Virtual Humans*, pages 77–108, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[34] Schrödinger E. *What is Life?: The Physical Aspect of the Living Cell; with Mind and Matter & Autobiographical Sketches*. Cambridge University Press, 1967.

[35] Tegmark M. Consciousness as a state of matter. *Chaos, Solitons & Fractals*, 76:238 – 270, 2015. URL: http://www.sciencedirect.com/science/article/pii/S0960077915000958, doi:https://doi.org/10.1016/j.chaos.2015.03.014.

[36] Tononi G. An information integration theory of consciousness. *BMC Neuroscience*, 5(1):42, Nov 2004. doi:10.1186/1471-2202-5-42.

[37] Tononi G. and Koch C. Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 370(1668), 2015. URL: `http://rstb.royalsocietypublishing.org/content/370/1668/20140167`, `arXiv:http://rstb.royalsocietypublishing.org/content/370/1668/20140167.full.pdf`, `doi:10.1098/rstb.2014.0167`.

[38] Velmans P. M. How to define consciousness – and how not to define consciousness. *Journal of Consciousness Studies*, 16(5):139–156, 2009.

[39] Wittgenstein L. Tractatus logico-philosophicus. 1922.

[40] Wittgenstein L. *Philosophical Investigations*. Basil Blackwell, Oxford, 1953.

[41] Zahavi D. Killing the straw man: Dennett and phenomenology. *Phenomenology and the Cognitive Sciences*, 6(1-2):21–43, 2007.

Received 8.10.2018, Accepted 21.11.2018