Minds and Machines (2018) 28:515–541
https://doi.org/10.1007/s11023-018-9468-3

From Computer Metaphor to Computational Modeling: The Evolution of Computationalism

Marcin Miłkowski[1]

Received: 25 September 2017 / Accepted: 26 June 2018 / Published online: 2 July 2018

Abstract
In this paper, I argue that computationalism is a progressive research tradition. Its metaphysical assumptions are that nervous systems are computational, and that information processing is necessary for cognition to occur. First, the primary reasons why information processing should explain cognition are reviewed. Then I argue that early formulations of these reasons are outdated. However, by relying on the mechanistic account of physical computation, they can be recast in a compelling way. Next, I contrast two computational models of working memory to show how modeling has progressed over the years. The methodological assumptions of new modeling work are best understood in the mechanistic framework, which is evidenced by the way in which models are empirically validated. Moreover, the methodological and theoretical progress in computational neuroscience vindicates the new mechanistic approach to explanation, which, at the same time, justifies the best practices of computational modeling. Overall, computational modeling is deservedly successful in cognitive (neuro)science. Its successes are related to deep conceptual connections between cognition and computation. Computationalism is not only here to stay, it becomes stronger every year.

Keywords Computationalism · Computational theory of mind · Computational neuroscience · Computer metaphor · Information processing · Computational modeling · Computational cognitive science · Working memory

1 Introduction

In this paper, I argue that computationalism is a progressive research tradition. No longer does it use mere computational metaphors. Instead, it offers empirically validated computational models, which can be best understood as offering mechanistic

✉ Marcin Miłkowski
marcin.milkowski@gmail.com

[1] Institute of Philosophy and Sociology, Polish Academy of Sciences, ul. Nowy Swiat 72, 00-330 Warsaw, Poland

explanations. The argument proceeds in the following fashion: First I argue that "the computational theory of mind" is a misnomer. It is not a theory but a rich, evolving research tradition. Then I review early conceptual arguments for the computational study of nervous systems. These arguments are no longer compelling, but can be restated in the light of the mechanistic account of physical computation. They also substantiate best research practices in computational neuroscience. To show this, I briefly show how research on working memory has progressed.

In particular, my focus is on validating computational models by empirical evidence, which, according to critics of computationalism, should not be at all possible.[1] The care for the empirical validity of models stems naturally from their methodological assumptions. These assumptions can be understood best in the light of the mechanistic approach to physical computation, rather than in terms of the semantic approach to computation, or three-leveled methodology of computational explanation introduced and developed by Marr (1982).

The paper is structured as follows: In the next section, I review early conceptual arguments for computational modeling in psychology and neuroscience. These arguments are not entirely compelling, but can be salvaged by turning to the mechanistic theory of explanation (Sect. 3). In Sect. 4, I turn to two cases, and argue that there has been considerable research progress in computational neuroscience, which shows that the mechanistic account is both descriptively and normatively accurate for computational neuroscience. The early models are illustrated here by the work of Young (1964, 1978) on memory, based on his groundbreaking research on cephalopods. They were largely speculative and metaphorical, and the key problem was how to evaluate computational models empirically. While Marr's systematic approach led to considerable progress in cognitive modeling, it did not solve this problem satisfactorily, and cognitive models have continued to be merely consistent with neuroscientific evidence. However, contemporary modeling methodology goes beyond mere consistency, and is poised to offer new predictions. This is illustrated by the work of Eliasmith et al. (2012) on the large-scale functional model of the brain. This kind of methodology is required by the mechanistic—but not semantic—approach to computational explanation. The fact that current best practices are in line with the mechanistic approach lends support both to this approach, and to computationalism.

2 Computationalism as a Research Tradition

While it is customary to use terms *the computational theory of mind* and *computationalism* interchangeably, this custom may be misleading. Theories are usually thought to be not only explanatory but also predictive. Computationalism remains too abstract to predict almost any non-trivial fact about biological brains. The main assumption of computationalism is that nervous systems are computers, and cognition can be explained in terms of nervous system computation. This assumption substantiates the claim that biological agents engage only in tractable computation

[1] A short survey of objections to computationalism is offered by (Miłkowski 2017a).

(Frixione 2001; Rooij et al. 2010). But it is not of much use when explaining phenomena which are typically listed in textbooks for cognitive psychology.

Computationalism is not a theory but a research program that is helpful in creating both theories and individual models of cognitive phenomena. While the notion of the research program has been introduced to the philosophy of science by Lakatos (1976) to account for scientific change and rationality, one part of his account does not fit the dynamics of research well. Lakatos assumes that there is an unchangeable core in any given research program. Laudan (1977) has shown that this is rarely the case. Programs usually change over time. Moreover, Laudan has stressed that the fruitfulness of research programs lies not only in predicting empirical facts, but also in theoretical problem solving, which has been ignored altogether by Lakatos.

To make it clear that Laudan's account is embraced here, his notion of *research tradition* will be used instead. He cites three characteristics of research traditions (Laudan 1977, pp. 78–79). First, "every research tradition has a number of specific theories which exemplify and partially constitute it." Indeed, there are multiple approaches to brain computation, from classical symbolic computation, to connectionism, to dynamic spiking neural networks. Second, they exhibit "certain *metaphysical* and *methodological* commitments which, as an ensemble, individuate the research tradition and distinguish it from others." I analyze these methodological and metaphysical commitments below in this section. Third, traditions go through a number of formulations and usually have a long history, which is easily evidenced, both in changes of basic assumptions, and in development of computational models of working memory, as described in Sect. 3. Early models were fairly sketchy and computational talk could be understood to be merely metaphorical. This is not the case anymore.

One could continue to talk of the computational theory of mind, while stressing that the term itself is an umbrella term. However, it is less misleading to think of computationalism as a diverse research tradition composed of multiple, historically variable computational theories of mind (or brain). By conflating the research tradition with one of the early theories, one could be tempted to reject the whole tradition. In particular, one might conflate computationalism with one of its versions, usually dubbed "GOFAI," or Good Old-Fashioned Artificial Intelligence. John Haugeland defined GOFAI as the assumption that intelligence is based on the capacity to think reasonably, and reasonable thinking amounts to a faculty for internal "automatic" symbol manipulation (Haugeland 1985, p. 113). Haugeland's formulation is thought to be related, for example by Margaret A. Boden (2014), to the physical symbol system hypothesis defended by Allen Newell and Herbert Simon (Newell 1980; Newell and Simon 1976).

The physical symbol system hypothesis states that a "physical symbol system has the necessary and sufficient means for general intelligent action" (Newell and Simon 1976, p. 116). The problem with taking this claim at face value would be that even universal computers may run extremely simple software throughout their lifetime. They may have precious little to do with what we may call "general intelligent action," in particular because some of them may take no inputs at all. This is also why Newell and Simon elucidate what they mean by "sufficient means": "any physical symbol system of sufficient size can be organized further to exhibit general

intelligence," (Newell and Simon 1976, p. 116). In other words, while it may seem that computationalism claims that cognition is any computation, the proponents of this idea did not mean that *all* computation exhibits cognitive features; only some physical computation exhibits general intelligence. Thus, it is not just internal symbol manipulation; this manipulation has to be—in some sense—organized further.

The motivation for computationalism can be found in a number of texts on computer simulation in psychology. For example, Apter (1970, pp. 17–18), an early defender of the simulation approach taken by Newell and Simon, listed four ways that computers resemble brains:

(1) Both are *general-purpose* devices.
(2) Both are *information-processing* devices.
(3) Both may incorporate *models* within themselves.
(4) Both achieve their intellectual excellence by carrying out *large numbers* of *primitive operations*.

Let me, therefore, analyze the ways in which these properties are understood by Apter to show that the analogy has to be rephrased in modern terms. Then I will also briefly analyze influential methodological advice that has shaped computational modeling in psychology and neuroscience.

2.1 Flexibility

Both computers and brains are general-purpose devices: they can operate in a variety of ways, which makes them adaptable to their environment, and this, in turn, can explain intelligence. This is exactly why Newell and Simon made *universal* computation part of their hypothesis. Newell's rendering of the argument deserves some attention:

> Central to universality is flexibility of behavior. However, it is not enough just to produce any output behavior; the behavior must be responsive to the inputs. Thus, a universal machine is one that can produce an arbitrary input–output function; that is, that can produce any dependence of output on input (Newell 1980, p. 147).[2]

But are cognitive agents actually *unboundedly* flexible in their behavior? Is there evidence that, say, human beings are able to adapt to *any* environmental constraints? To demonstrate this empirically, one would have to actually observe an infinite

[2] One could argue that Newell actually uses a notion of universality that is different than the one used in the computability theory, as he proceeds to talk of behavioral flexibility in spite of limited physical resources (I thank one of the anonymous reviewers of this paper for this comment). However, the physical symbol system is then proven to be computationally equivalent to a universal Turing machine. My claim here can be understood in line with Newell and Simon's idea that one has to understand intelligent action as constrained and enabled by the structure of the environment; therefore, impressive behavioral complexity and flexibility can be a feature of a relatively simple organism (Simon 1956). Rationality can occur without universal computation.

number of behaviors displayed by human beings, which is, alas, physically impossible for finite observers. The only evidence we have is that people actually sometimes learn to adapt, and to control their environment. But sometimes they fall short of recognizing more complex causal relationships, because they rely too much on the received feedback (for a discussion and review of experiments, see Osman 2014). In other words, the hypothesis that behavior is infinitely flexible not only cannot be confirmed, but also seems to be disconfirmed, This is shown by a number of phenomena related to our illusions of control or illusions of chaos, as well as by highly stereotypical behavior triggered by situational factors (Wood and Neal 2007). Human behavior may be *vastly* flexible but not *infinitely* flexible. This is even more true of animal behavior, which is sometimes highly stereotypical.

However, no physical computers ever literally *instantiate* a universal Turing machine. We only idealize finite-state machines as universal Turing machines for the sake of simplicity. As Gary Drescher states: "There are no precise rules governing the suitability of this idealization; roughly, the idealization is appropriate when a finite-state automaton has a large array of state elements that it uses more or less uniformly that thereby serve as general memory" (Drescher 1991, p. 38). PCs, smartphones, and tablets are all, in reality, just finite-state automata, with the ability to compute a large class of functions that a universal Turing machine also computes. However, they do not compute all of them: their memory or input is not really unbounded, hence, they are not physical universal machines. However, these finite-state machines can be made quite flexible by running various kinds of routines in a given situation.

Newell admits that real systems are bounded in their resources, but retorts that "the structural requirements for universality are not dependent on unbounded memory, only whether the absolute maximal class of input–output functions can be realized" (Newell 1980, p. 178). What counts is not unbounded, but "sufficiently open memory" and "the structure of the system," because for "every machine with unbounded memory, there are machines with identical structure, but bounded memory, that behave in an identical fashion on all environments (or problems) below a certain size or complexity" (Newell 1980, p. 161). However, from the mathematical point of view, these machines are not technically universal, even if they may be usefully idealized as such.[3]

Other theorists were also at pains to show that cognition cannot happen without universal computation. For example, Pylyshyn (1984) claimed that cognition requires not only flexibility, which in turn requires universality, but also that semantic interpretability requires universality. In other words, he has claimed that finite-state machines (FSM) cannot be semantically interpretable; however, this argument has been rebutted by Nelson (1987). Moreover, Pylyshyn relies on the argument from natural language acquisition, which claims that natural languages require universal computation (Pylyshyn 1973, pp. 26, 1984, pp. 87–88). The argument is stated thus: at least some natural languages that human beings acquire are so

[3] It is trivially easy to show they are not universal by proving a pumping lemma for a given bounded-memory machine.

complex that they require a universal Turing machine to process them (technically speaking, they are context-free languages). However, the empirical evidence for the claim is ambiguous and scarce (Shieber 1985).

Additionally, universal computation may seem brittle in some environments, lacking complex emotional motivation and being "single-minded," which has been one of the major worries of critics of computationalism for years (Neisser 1963). This brittleness is now informally referred to as the inability to solve the frame problem (Dennett 1984; Wheeler 2005). Universal computation does not solve the brittleness problem by itself (even if the frame problem in the proper technical sense of the term is, in principle, solvable; cf. (Shanahan 1997). As Newell and Simon would probably repeat, this computation has to be "further organized."

To sum up, the arguments that were supposed to justify the explication of "general-purpose computation" in terms of universality were unsuccessful. There is no empirical evidence that, say, human memory is actually unbounded, or that it could be unbounded if human beings were to use external resources such as pencil and paper [as suggested by Turing (1937) in his initial idealization of human computers]. Universal computation is neither necessary nor sufficient for behavioral flexibility.

2.2 Information-Processing

The second aspect of the analogy is that both computers and minds (or brains) process information. Some contemporary theorists, in particular those inspired by the ecological psychology of James J. Gibson, reject the notion that people process information just like computers (Chemero 2003). But their criticism is related more to the issue of representational information (see Sect. 2.3). The same is at stake with regards to the enactive account of cognition. Enactivists have claimed that autonomous biological systems are all cognitive, but that cognition does not involve any information intake: "The notions of acquisition of representations of the environment or of acquisition of information about the environment in relation to learning, do not represent any aspect of the operation of the nervous system" (Maturana and Varela 1980, p. 133).

However, the information at stake need not be semantic information, if it were, this part of the analogy would be included already in the capacity for incorporating internal models. The notion of information sufficient to express this aspect of the analogy is *structural information content*, as defined by MacKay (1969), which indicates the minimum equivalent number of independent features which must be specified, i.e., the degrees of freedom or logical dimensionality of the information vehicle (whether or not it represents something). This kind of information is pervasive, and present in any causal process (Collier 1999). One can speak of structural information content in living beings as well.

To sum up, it is not controversial to say that brains and computers process structural information, which may, or may not, be digital. It is much more controversial, however, to say that nervous systems manipulate internal symbols that represent reality.

2.3 Modeling Reality

The third aspect of the analogy between brains and computers is that both can create *models* of reality. In other words, one of the basic functions of the brain is to build models of the environment. Apter (1970, p. 18) points out that this was particularly stressed by J.Z. Young, whose work is the focus of the next section on working memory, and by Kenneth Craik (1943), one of the early proponents of the structural representations:

> By a model we thus mean any physical or chemical system which has a similar relation-structure to that of the process it imitates. By 'relation-structure' I do not mean some obscure non-physical entity which attends the model, but the fact that it is a physical working model which works in the same way as the process it parallels, in the aspects under consideration at any moment. Thus, the model need not resemble the real object pictorially; Kelvin's tide-predictor, which consists of a number of pulleys on levers, does not resemble tide in appearance, but it works in the same way in certain essential respects (Craik 1943, p. 51)

Craik's approach is similar to what is now called *structural representations* (Cummins 1996; Gładziejewski and Miłkowski 2017; O'Brien and Opie 1999; Ramsey 2007). However, not all proponents of computational approaches to the study of cognition embrace the structural account of representations.

Notably, many computationalists rely on the notion of *symbol*, whose most important feature (in contrast to structural representations) is that the relationship between its vehicle and its content is arbitrary (Fodor 1975; Haugeland 1985; Newell 1980; Simon 1993). The notion of the symbol used by some of these theorists, however, remains ambiguous (Dennett 1993; Steels 2008). Sometimes Newell and Simon talked of symbols in the sense of the computability theory, where they are just distinguishable entities in a computing device (e.g., on a Turing machine tape). Sometimes they thought of Lisp symbols, which are pointers to other data structures, and sometimes as elements that designate or denote elements of reality (for further analysis and examples, see Miłkowski 2016a). While it is understandable that their intention was to show that computers can operate not only on numbers, but on arbitrary elements, and that the obvious term for such an arbitrary element is *symbol*, their terminological confusion makes it difficult to evaluate what they meant by *symbol* in any context.

At the same time, symbolic and structural representations are thought to be opposite, because the vehicles of structural representations are not *arbitrarily* related to their contents (Sloman 1978). Suppose that computation is manipulation of internal symbols. Is processing structural models not computation? The notion of symbol is too ambiguous to provide a single answer. It is far from clear whether even the broadest notion of the symbol, understood as a distinct part of the Turing machine's alphabet, could be applicable to analog computational models, which were prominent among early theorists of artificial intelligence and computer models in neuroscience (Von Neumann 1958; Young 1964).

Moreover, unclear ideas about intentionality or aboutness, analyzed in terms of syntactical relationships between symbols by many proponents of symbolic computation (Haugeland 1985), might engender the appearance that computationalists confuse syntax with semantics (Harnad 1990; Searle 1980). The computationalist may say that one can treat syntax as semantics in an idealizing fashion (Dennett 1987; Haugeland 1985). But this idealization is not forced by computationalism itself. Craik and Young had little reason to think of their models as arbitrary pieces of syntax.

This part of the early computationalism is in need of overhaul and reappraisal.

2.4 Complex Behavior Out of Simple Building Blocks

The fourth part of the analogy is that both computers and brains have a large number of similar building blocks that perform simple operations, which, taken together, contribute to a complex behavior. This was part of the first computational model of the nervous system, which modeled the system as a large collection of logic gates (McCulloch and Pitts 1943). It has remained a part of the connectionist research program, which underlined the point that the connections between (computational) neurons count the most, not individual neurons.

However, it is not a conceptual truth that individual neurons may perform only simple computations. Some argue that single neurons may be suited to efficient computation of tree-like data structures (Fitch 2014). If this is true, it does not mean that computationalism came to a grinding halt. It just shows that it is not only connections between neurons that count in neural computation. Therefore, we need to revise this part of the analogy as well.

2.5 From Analogy to Strict Algorithms

The notion that mere analogy is not enough was a major methodological breakthrough in computational neuroscience—the function to be modeled should be described in a strict fashion. This influential methodological advice for the computational study of cognition was formulated by David Marr (1982) in his three-leveled[4] account of computational explanation.

At the first level, called *computational*, the explanations deal with operations performed by the system, and the reason it performs them. In other words, one assumes that the system in question is flexibly adapted to some environment (see Sect 2.1), in which some kind of input–output transformations are useful (see Sect 2.2). At the

[4] Marr's levels of explanation are not literally *levels* because they are not hierarchically ordered by any clear relationship, be it composition, constitution, realization, implementation or abstraction. Notice, for example, that the computational level includes the consideration of the issue of why certain computation is executed in a certain environment, but neither the adaptation to the environment nor the environment itself is implemented by algorithms and representations. Marr's levels are, rather, aspects of all computational systems that Marr considers to be jointly necessary and sufficient, in order to explain their computational processes.

second level, one specifies the representation for the input and output, and the algorithm for the transformation of the input into the output. Although Marr's approach to representation may be interpreted variously, it clearly mirrors the assumption that cognitive computational systems model their worlds (Sect. 2.3), and that they perform certain basic operations (Sect 2.4). These basic operations and representations must be shown to be biologically realized at the last Marrian level, the level of implementation. Only then is the explanation complete.

The popularity of Marr's approach should be no surprise, given that it makes use of the basic aspects of the general analogy involving the computer and the brain. By following Marr strictly, one can pursue the goal of the computational research tradition, and offer detailed computational models for cognitive phenomena. Marr's methodology, quite obviously, is theoretically almost vacuous as it comes to cognition, and remains very abstract. Almost any of the competing computational models in cognitive science can be cast in terms of the three levels. This methodology offers only a check on whether all the important questions have been asked. But there are no preferred answers.

In spite of the popularity of Marr's methodology, it is rarely observed that Marr could not make sense of some computational approaches. For example, he criticized the heuristic approach taken by Simon and Newell as mere gimmickry (Marr 1982, pp. 347–348). According to Marr, when people do mental arithmetic, they actually do something more basic. Here, Marr seems to be committed to a version of the basic operations assumption which states that modeling should refer to the *really* basic building blocks. But his own account of levels cannot justify this commitment; nowhere does he justify the claim that one should specify mental arithmetic in terms of the operations of neural ensembles, or of individual neurons, for example. This is not a coincidence, his account of the implementation level remains awkwardly abstract, which makes it the Achilles heel of his methodology.

To summarize this section, the early statements of basic theoretical assumptions are faulty in several ways. They unpack the analogy between nervous systems and computational devices in an outdated way. Marr's level-based methodology was a serious improvement, but suffered from the problem of how to include the evidence about the work of the nervous system. In the next section, I show that the analogy can be motivated better from the contemporary mechanistic point of view. In particular, the mechanistic solution sheds light on the role of neural evidence.

3 Mechanistic Computationalism

In this section, I defend the claim that the basic metaphysical assumptions of the computationalist research tradition are best understood from the mechanistic point of view. The mechanistic computationalism is compatible with a variety of approaches to computational modelling without deciding contentious matters of fact a priori, such as whether there are genuine representations in all computational systems or not. All it requires is that explanatory factors be actually explanatorily relevant and in line with empirical evidence. This is what drives the development of methodology in computational neuroscience. After briefly elucidating the notion that nervous

systems are computers from the mechanistic perspective, I rephrase the basic metaphysical and methodological assumptions reviewed in the previous section. Then, in Sect. 4, I illustrate the development of methodology by showing that empirical validation of models is guided by the mechanistic norms of explanation.

The mechanistic account of physical computation holds that computers are physically organized systems, whose component parts and operations jointly produce computational phenomena as a result of containing information vehicles[5] (Miłkowski 2013; Piccinini 2015). These vehicles need not carry digital information; at least conceptually, there may also be information vehicles that require an infinite number of degrees of freedom in order to be specified precisely.[6] All physically implemented computation requires transformation of information vehicles (in the extreme case, the transformation at issue may be identity). Piccinini (2015, p. 121) requires that this transformation occurs according to rules sensitive only to properties of vehicles of information, while rules are understood as mappings from input vehicle properties (and possibly some internal states) to output vehicle states. In extreme cases, the inputs of a computer are empty.

At least some operations of the nervous system necessarily involve information processing in this sense. The hypothesis that nervous signals, both electrical (such as spikes) and chemical (the release and uptake of neurotransmitters) are informational, is at the core of current neuroscience. This is how the mechanistic account restates the second part of the analogy from the previous section, which was the least problematic.

3.1 Bottoming Out

The operations and component parts of computational mechanisms have to be explained, at some point, in a purely causal manner, or to *bottom out* in non-computational elements (cf. Machamer et al. 2000; Miłkowski 2012). In a sense, the mechanistic view treats the term *physical* as used in the expression *physical symbol system hypothesis* somewhat more seriously than Newell and Simon. But it does not force bottoming out at the level of elementary physics. The level is decided pragmatically by the research community.

It follows that there may be some features of cognition that cannot be explained by reference exclusively to their computational structure. But this is not a surprise. For example, standard laptop computers generate heat, but this cannot be explained solely by appealing to the software that runs on them. The physical properties of electronic parts, in particular CPUs, are responsible for heat dispersion when electricity runs through them. Similarly, cognitive processing is studied by observing

[5] To be exact, Piccinini eschews the notion of information and talks of medium-independent digits, but then concedes that they need not be digital. This is a purely verbal difference.

[6] The caveat "at least conceptually" is intended to make it clear that it may be impossible for physical reasons. Some might argue that any measurement operation on physical information vehicles has finite resolution, and some error is involved. Or one could deny that there are actual infinities out there. This paper remains neutral in this regard.

response times of experimental subjects (Meyer et al. 1988; Posner 2005). Response times cannot be fully analyzed without considering the inherent hardware speed (usually assumed to be a constant multiplier). Thus, the speed of the nervous processing is not a solely computational fact. It is a feature of the physical implementation of computation (Miłkowski 2011).

The bottoming-out principle is related also to the basic operations stressed by Marr. Basic operations are important, because they are part of testing the validity of computational models. Validation of *implemented* computational models with experimental data has become indispensable when publishing serious work in cognitive science journals.

One early important requirement was formulated by Jerry Fodor in his first book on psychological explanation (Fodor 1968). He distinguished two kinds of relationships between computational simulations and experimental data. First, models can be *weakly equivalent,* as long as they produce the same output, given the same input. These models, however, as Fodor stressed, are not satisfactory. Only the second equivalence class, of *strong equivalence*, offers a full explanation. A model is strongly equivalent when it is not only weakly equivalent, but also when "the processes upon which the behavior of the machine is contingent are of the same type as the processes upon which the behavior of the organism are contingent" (Fodor 1968, p. 138).

But how does one establish that the types of processes are the same as each other? Without sophisticated measuring instruments, computational models can never be shown to be strongly equivalent. In spite of the precision with which we can state these models, the differences between them could never be resolved empirically.

This became even more pressing when proponents of computationalism also embraced the claim that psychological explanations are autonomous from neuroscientific evidence (Fodor 1974). Such autonomy comes at the price of empirical indeterminacy. John Anderson argued that the debates between various approaches to mental imagery, for example, could not be decided by recourse to behavioral data. In addition, physiological evidence was rather indirect and too theoretically mediated to be of any use (Anderson 1978). His argument also relied on the observation that the debate failed to specify the processes for handling diverging formats of mental representation. This, he stressed, might deliver the constraints required in order to go beyond behavioral data. The subsequent research, rather unsurprisingly, included neuropsychological data (Farah 1988).

Today, various kinds of neuroscientific evidence are significant for computational models, from brain imaging (e.g., Anderson et al. 2008), experimental interventions such as transcranial magnetic stimulation (Pascual-Leone et al. 2000) and natural brain damage, to patch clamps, which are used to record electric signals from neurons directly, and optogenetic methods in model animals (Deng et al. 2017). All in all, although a single piece of neuroscientific evidence is usually not univocal, it helps to constrain the space of plausible computational models of cognition (Bechtel 2008), because validation usually uses multiple pieces of evidence to further constrain this space. For example, in the case of computational psychiatry, a wide range of various kinds of evidence is already used for model validation (cf. Bhattacharya and Chowdhury 2015).

One particularly important requirement has been identified in a recent discussion (Coltheart 2013; Mole and Klein 2010). It was noted that a large number of cognitive models were supposed to be valid owing to their consistency with neural data. But this is not enough to establish validity. Validation should always be contrastive; not only should a model be shown to be better than alternative models, but it should also have predictions about neural data that could be used to disconfirm the model (or to refine it).

3.2 General-Purpose Computational Mechanisms

Let us now turn to the idea of general-purpose devices as underlying cognitive flexibility. At least in some areas of comparative psychology, flexibility is indeed thought to be an important distinctive feature of cognitive processing (Buckner 2015). But flexibility can be explained by appealing to non-universal computation. What is more important for behavioral flexibility is to see whether computational mechanisms may solve behavioral problems to allow the agent to deal with its environment effectively. To be a behaviorally flexible organism, it's sufficient to respond to a large class of inputs by producing a large class of behaviors. For example, it seems biologically plausible that evolutionarily early nervous systems, or their predecessors, are computationally weaker than universal Turing machines, even if one could argue that human brains are helpfully idealized as universal Turing machines. Thus, what is more crucial to flexibility is the kind of cognitive architecture rather than the ability to implement universal computation.

Non-universal computation is still computation. Therefore, the mechanistic interpretation of the basic computationalist assumption makes room for non-universal computation. To be general-purpose, computational mechanisms need not be universal machines.

The mechanistic perspective makes it possible to consider also non-Turing computation as part of the computationalist research tradition. The account does not decide which mathematical model of computation is viable and which is physically possible. In other words, it does not quickly presuppose the Church-Turing thesis, which states that the class of computable functions is equivalent with what a universal Turing machine may compute [the relation of this thesis to physical computability is analyzed further by Piccinini (2011)]. Simply speaking, philosophers cannot decide, in the confines of their armchairs, which mathematical notions of computation are acceptable. The question of whether these notions are also physically possible or actual may require extensive empirical inquiry in the realm of physics (and maybe also computer engineering). Therefore, the issue of what is contained in the class of (physically) computable function is left open, or transparent (Chrisley 2000). Philosophical accounts should not presuppose what may happen in the course of the future development of mathematics and computer science, they should remain as open as possible. In other words, there may still be some room for development in the notion of computation; it may still have some open texture (Shapiro 2006). At the same time, the accounts of physical computation should be able to spell out

precisely what is required to implement, say, digital computers, or some type of analog computation.

This is all the more important because it is plausible that biological brains are not necessarily digital computers—some of their computation may be analog, or even partially digital and partially analog (Maley 2017; Piccinini and Bahar 2013). Analog computation, as well as hypercomputation, is computation at least in the sense that they are both the focus of mathematical and engineering theories of computation. Nobody would be fired from a computer science department for axiomatizing analog computation or hypercomputation. Both are still in purview of computer science.

Therefore, according to the mechanistic perspective, computation is information-processing in the non-semantic sense, and general-purposiveness is not to be equated with universal Turing computation. Moreover, elementary processing components may be also fairly complex in computational terms; thus, if Fitch is right that neurons are particularly optimized to perform computations on tree data structures, computationalism does not fail.

3.3 Neural Representations

The most difficult problem with traditional assumptions of computationalism is how to precisely understand that both brains and computers "incorporate models." How are these models related to symbols cited in the physical symbol system hypothesis, or to internal models assumed by GOFAI?

Remember that computationalism is a broad research tradition, so one should expect controversy around important issues. However, there need not be one computational way of thinking about representations. Overall, the issue of what makes a cognitive representation is complex. The ideas of structural representations (or models) and interpreted symbols are not the only ones in the game. Others relied on the notion of *indication* (Dretske 1986; Fodor 1984), and some consider representational talk a mere *gloss* on computational explanations (Chomsky 1995; Egan 2010). From a mere gloss, it's just a short stretch to instrumentalism (Dennett 1971, but see his 1991 for a more realist position), fictionalism (Sprevak 2013), eliminativism (Downey 2017), or general anti-realism about mental content (Garzon 2008; Hutto and Myin 2013).

Thus, it would be a mistake to think that there is one approach to representation defended by *all* the proponents of computationalism. One of the advantages of computationalism is that it can fruitfully explicate some features of representation and fits diverse approaches to the aboutness of representations.

First, computationalism can make content causally relevant, thus helping to solve one of the facets of the mind–body problem. Because physical vehicles are causally relevant in computation, mental representations can be causally efficacious, too (Newell 1980). However, only the syntactical features, many people are quick to add, are causally relevant. According to Dennett, semantic engines are, strictly speaking, impossible, while what is at stake are syntactic engines (Dennett 1987). This argument is embraced by many proponents of

anti-representationalism. However, as Gładziejewski and Miłkowski (2017) recently argued, in the case of structural representations, it's not the structure of the vehicle *alone* that is causally relevant for the explanation of behavior. What is relevant in this case is rather the similarity relationship between the vehicle and the target of the representation. Therefore, semantic engines are not only logically possible, but, as they show by appealing to cognitive maps in rodents, also actual. To sum up, by assuming both computationalism and representationalism, one can explain how representational processes are causally responsible for behavior. There is currently no other naturalistic alternative for explaining how the contents of thoughts might cause action.

Second, by treating brain mechanisms as representational, we can fruitfully account for cognitive processes (Shagrir 2010a). At the most abstract level, representational explanations make it clear how cognitive systems may flexibly adapt to their changing environment; they are able to learn and predict it simply by building representations of it. While one could describe learning and adaptation in non-representational terms, the efficacy of control systems, in general, depends on their ability to represent the controlled domain (Conant and Ashby 1970). The attempt to re-describe them is mere wordplay.

To sum up, computationalism need not imply any particular version of representationalism, but it is a natural counterpart of the latter. Similarly, the mechanistic approach to physical computation states that computation is not *necessarily* semantic (Fresco 2010; Miłkowski 2013; Piccinini 2008). Thus, mechanistic computationalism is compatible with virtually all naturalistic approaches to mental content, including anti-representational computationalism (Villalobos and Dewhurst 2017).

All in all, the mechanistic proposal is to understand computationalism in a more open fashion. The argument that philosophers should not decide empirical issues applies to all parts of the analogy between nervous systems and computers: in transparent computationalism; in arguing that flexible action of agents need not rely on universal computation; in making the notion of information purely structural; and in openness toward various approaches to mental representation. Similarly, just like Newell and Simon, who did not propose any substantial theory of what cognition is, operationalizing it only in terms of general intelligent action, the mechanistic approach does not have a theory of what constitutes the mark of the cognitive (cf. Adams and Aizawa 2008; Rowlands 2009). This is a separate issue and it remains open (Akagi 2017). All that the mechanistic approach says is that computation is necessary for cognition, because cognition does not occur without processing information (in the minimal structural sense). Instead of claiming that it is just "further organization" of computation that would be sufficient to make a computational mechanism cognitive, mechanists also leave this question open to be decided by particular theorists. Moreover, the mechanistic approach admits that there could be facts about the mind or the brain that cannot be explained computationally.

Thus, in the mechanistic rendering of computationalism, its foundational assumptions may be shared by a variety of conflicting approaches. It does not conflate these foundational assumptions with any particular computational approach.

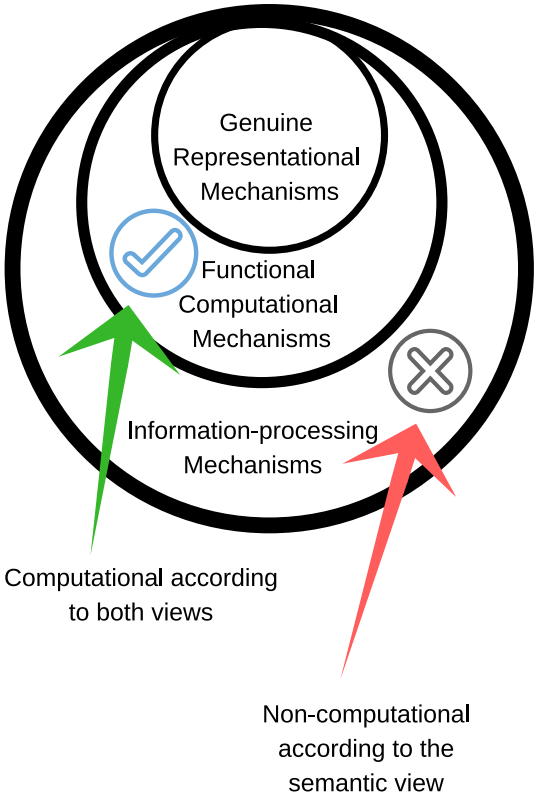

Fig. 1 Semantic versus mechanistic approach to physical computation. Both approaches agree as to the computational nature of functional computational mechanisms (marked by a checkmark), but the semantic account denies that generic information-processing mechanisms are computational (hence the X mark)

It is instructive to compare the mechanistic view of computation with the semantic approach (see Fig. 1), its greatest contender.[7] According to the semantic view, a spatiotemporal system is a computer if, and only if, some states of the system can be described as transforming some entities whose role is to model, or represent something (Fodor 1975; Shagrir 2010b). The semantic view is usually expressed in terms that seem to make computation a matter of ascriptions of representation (Dresner 2010). In contrast, the defenders of the mechanistic view claim that computation is an objective feature of mechanisms. Moreover, they differentiate between generic information-processing mechanisms, and mechanisms whose function is computation (Miłkowski 2013). For example, physicists may inquire about

[7] This is not to say that these are the only accounts ever proposed in the current debate (cf. Miłkowski 2013; for a review, see Piccinini 2015). One influential contribution is David Chalmers's (2011) causal account of physical computation (I thank the anonymous reviewer for flagging this omission in the previous version of this paper). However, because Chalmers does not rely on any particular account of causation nor causal explanation, his account suffers from a number of weaknesses from the mechanistic point of view. In particular, it does not appreciate the role of non-entirely computational features, such as response times, in the study of cognition (Miłkowski 2011). Other prominent views rely on the idea that a computer is just any physical system on whose states one can map computational states but suffer from indeterminacy problems (Putnam 1991; Searle 1992); cf. Sect. 5.2 below.

the ultimate computational limitations of the whole universe (Lloyd 2002), and this kind of question is not in any way unscientific. In contrast to generic information-processing mechanisms, computers, in the proper sense, are only those mechanisms whose function is to compute (Piccinini 2015). However, even these computers may not feature any genuine cognitive representations. Therefore, the class of representational computers is a subclass of functional computational mechanisms. Note: for the semantic view, all functional computational mechanisms are semantic in a certain minimal sense; specifically, they can be interpreted as such. So, the crucial difference between the semantic and mechanistic view, as the check mark on Fig. 1 shows, boils down to the class of non-functional information-processing mechanisms.

Importantly, however, the mechanistic view relies on a mature theory of explanation (Craver 2007; Glennan 2017; Machamer et al. 2000), while the semantic view is noncommittal as to the theory of explanation. It does not offer, by itself, methodological advice that is as rich the mechanistic approach. At the same time, the mechanistic approach can the insight of the semantic view that an important function of the nervous system is to model reality. Thus, the claim that the semantic and mechanistic view are mutually exclusive would be a false dichotomy (Miłkowski 2017b).

3.4 Computationalism and its Falsifiable Metaphysical Assumptions

Even if this mechanistic rendering of computationalism is relatively weak, that does not mean it could not be shown to be false. Simply, were information-processing causally irrelevant to cognitive phenomena, this approach would be bankrupt. It is unlikely this will ever happen. Computationalism has become the basis of a fruitful research tradition. In this section, the theoretical and metaphysical assumptions of the tradition were analyzed.

Research traditions evolve over time. That includes their methodological assumptions, in particular, the testing of models. In the next section, a historical and contemporary model of working memory is reviewed. I argue that the mechanistic approach to explanation vindicates contemporary computational neuroscience as offering genuine explanations. Computational neuroscience is no longer satisfied with metaphors. Its results are becoming more and more testable empirically.

4 How Neuroscience Studies Brain Computation

Early work in computational neuroscience consisted mostly of exploratory models, whose purpose was to show that, in principle, computational simulations suffice to produce interesting cognitive phenomena. The problem was how to build models that could be validated using available evidence. This can be illustrated in the work of J.Z. Young (1907–1997), one of the most influential biologists of the twentieth century (Boycott 1998). In the 1930s, one of Young's influential achievements was the development of study methods for spikes in giant nerve fibers of the squid,

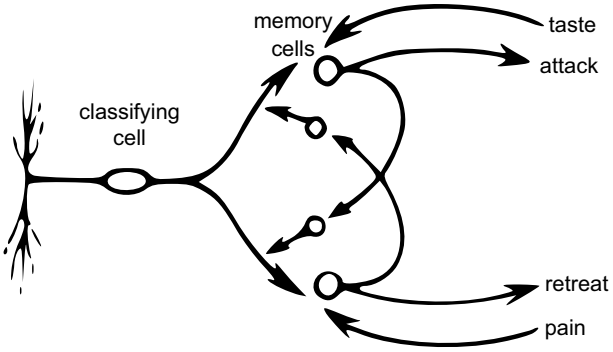

Fig. 2 Young's model of memory units in the octopus: Feature detectors cause changes in memory cells, which are coupled with taste receptors and attack effectors, as well as pain receptors and escape effectors. The arrows on the diagram are not fully interpreted, and their operation is merely localized in some structures of the nervous system. From Young (1965), by permission of Royal Society Publishing

which later led to groundbreaking results in modeling the electrical activity of the nervous system.

4.1 Memory Units in Cephalopods

Young's work was heavily influenced by cybernetics. His *A Model of the Brain* (1964) was one of the first attempts to bring together computational modeling and experimental work in neurobiology. Most computational psychology at the time remained silent about the nervous system. Young described the brain as the homeostatic computer involved in the control of behavior. He also discussed a number of experiments, which led him to distinguish two separate memory systems in the octopus: visual and tactile (Young 1965). One function of memory systems is to control behavior, or to change the animal's readiness to act in a certain way. While Young's hypotheses about the function and structure of learning mechanisms in the optic lobe of the octopus are heavily influenced by computer terminology, there are merely box and arrow diagrams of plausible mechanisms, with no accompanying computer simulations (see Fig. 2 for an example). Indeed, Young does not go into any detail of brain computation, merely pointing to simulations of analog computation used for pattern recognition (Taylor 1964).

Therefore, Young's work in the 1960s is indeed a prime example of what a computer metaphor might be, in biological research into the nervous system. He simply uses analogies from computer technology as his heuristics of discovery. Nonetheless, he does not defend any detailed computational model of the octopus's optical lobe. This drives his research on functional anatomy. For these purposes, analogies seem sufficient. However, a critic of computationalism might plausibly say that Young's functional anatomy has not led to mature computational neuroscience models, and rests upon loose talk.

Over the years, however, Young's model progressed. In the 1970s, in particular in his Gifford Lectures (Young 1978), instead of framing memory in terms of static

models or representations, he uses the term "program." He stresses at the same time that it is not meant literally, but expresses a concept of hierarchical and dynamical control. Overall, his use of computer models is best summarized as follows:

> Computers are in a sense artificial brains. Can we frame the descriptions that we want of brain function in the language used for computers? Computers are as it were a sort of new experimental animal that we can use for our studies, just as we use *Escherichia coli* or the giant nerve fibres of a squid. Clearly we are not made of transistors and it is the *principles* of computer software that we should consider. The analogy of computer programs can indeed be helpful—but only in the most general sense. Since we do not know how information is stored in the brain we cannot expect to use the detailed analogies of addressing and storing information, which is probably not done on the same principles in brains and computers (…). What can be used is the concept of hierarchy of controls (Young 1978, p. 62).

In the Gifford Lectures, "programs of the brain" are related to cyclical operations of the nervous system, as driven by physiological rhythms. They are hypothesized to be a result of the selection (trimming) of nervous connections, and they play the role of selecting behaviors, as based, for example, on emotional processing and learning. However, Young's efforts at explaining how learning occurs in terms of rewards (Young 1978, pp. 88–89) fall far behind mathematical models of reinforcement learning (Rescorla and Wagner 1972), such as the temporal difference algorithm (Sutton 1988). At the same time, Young's account of basic memory units—*mnemons* in his terminology—is not entirely loose, as the mnemon is supposed to "record one bit of information" (Young 1978, p. 87).

Once again, in the 1990s, Young returned to mnemons and framed the computational principles of two memory systems of the octopus in terms of distributed processing in "serial networks, with recurrent circuits" (Young 1991, p. 200). In spite of significant theoretical progress, the criticism against this kind of work may be repeated: the use of computation remains largely heuristic and does not yield precise experimental predictions. Moreover, neuroscientific evidence is not used explicitly to reject computational hypotheses. Indeed, these models of nervous mechanisms are largely schematic. Computational detail is lacking.

4.2 Semantic Pointer Architecture and Serial Working Memory

To make the contrast with J. Z. Young's model of memory dramatic, I focus here on the Semantic Pointer Architecture (SPA) recently proposed by Chris Eliasmith and his collaborators (Eliasmith 2013; Eliasmith et al. 2012). This model explains serial working memory. It goes beyond recurrently connected networks (to which Young also gestured in the 1990s). The SPA is modeled using spiking neuron networks. It has become the basis of the largest cognitive model of the brain to date. This is the feature of Eliasmith's approach to modeling that relies both on behavioral and physiological data—in this case, on psychological experiments related to serial recall (Baddeley 1998) and animal studies (although these are rare). Two fundamental

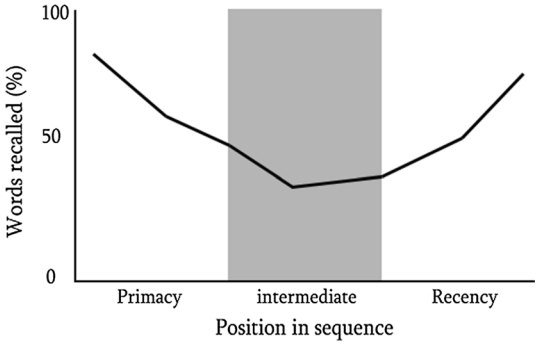

Fig. 3 Primacy and recency effects in serial recall of words. Source: From English language Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=61092398

features of serial working memory are observed: primacy and recency; first and last items are best recalled, which leads to a U-shaped curve of recall (see Fig. 3).

The model of ordinal serial encoding of memory proposed by Xuan Choo of Chris Eliasmith's lab is mapped onto cortical and hippocampal memory systems. It is implemented in an idealizing fashion. The input of the model is a semantic pointer (or a structure that compresses information from another part of the neural network in a lossy fashion) representing the item, and another pointer that represents its position. These are processed and added to memory. The "overall representation of the sequence is the sum of the output of the two memories" (Eliasmith 2013, p. 214). Consequently, this model has become part of the larger brain simulation, Spaun (SPA Unified Network), and underlies its performance in several cognitive tasks. Not only does it exhibit primacy and recency effects, but also matches averaged human data "from a rapid serial-memory task using digits and short presentation times (…) is also evident, with 17 of 22 human mean values within the 95% confidence interval of 40 instances of the model" (Eliasmith et al. 2012, p. 1204).

However, mere consistency is not enough. Other models can reproduce recency and primacy effects. In contrast, Spaun reproduces the psychophysical regularity known as Weber's law, which states that the variance in response time increases with the mean response time, as well as producing a behavioral prediction that a certain type of question posed to human subjects will not affect their accuracy in recall. More interestingly, Spaun also yields a neurophysiological prediction about the expected change of similarity in neural firing, in the dorsolateral prefrontal cortex during a serial working memory task. The dorsolateral prefrontal cortex is assumed to be the temporary storage and manipulation of higher level data related to cognitive control.

One striking feature of the working memory model in SPA is that its performance could only be assessed by analyzing a working computational simulation. The equations describing the model specification cannot reproduce the recency and primacy effects (Eliasmith 2013, p. 218). Only in the implemented model was it evident that "the individual neurons themselves saturate when participating in the representation of large vectors" (Eliasmith 2013, p. 219). As Eliasmith stresses, this demonstrates the theoretical significance of "constructing a neural implementation for explaining high-level psychological behavior," and goes against the received view

that psychological and neuroscientific theories are independent. Hence neural implementation is relevant to the explanation of behavioral data. This is exactly what was missing in J. Z. Young's work—no computer implementation was ever produced, which, from the mechanistic point of view, shows that this work is not as valuable as Eliasmith's (Miłkowski 2013, p. 133).

Moreover, Eliasmith's approach, in contrast to many connectionist strategies, is to build models directly by relying on theoretical considerations, instead of using machine learning techniques to set connection weights in artificial neural networks. To use Marrian terminology, one could say that algorithms and representation are not made obsolete by relying on the computer implementation. From the mechanistic point of view, the dynamic model of brain computation used by Eliasmith is indispensable in assessing whether the explanation is satisfactory. Simply, without it, one could not see whether the causal connections in the brain are reflected in the computation (Kaplan 2011). In other words, the mechanistic perspective does not obviate the need to understand the algorithmic principles of brain computation.

The case of the working memory model shows that the field of computational neuroscience has progressed from proposing simple "proof-of-concept" models, which were supposed to prove that a certain cognitive capacity might be realized computationally, as evident in Young's theorizing, to models with at least partial biological plausibility. The implicit assumption of this development is that biological plausibility or realism is explanatorily crucial. This, it may be argued, is how this research tradition has responded to various criticisms which stated, in a nutshell, that computational modeling is unbearably easy. One can create a plethora of very precise models without being able to say which one is approximately true. By providing appropriate evidence about brain operations and structures, one can substantiate the claim that the computational model is sufficiently developed at the level of implementation, as Marr would call it. Instead of ignoring the evidence as non-essential—which is indeed inessential when producing mere proof-of-concept models—modelers are busy validating their models with empirical data. Of course, models are usually strategically idealized, but this does not mean that they are unbearably metaphorical anymore.

4.3 Interim Summary: Research Tradition in Full Swing

Computational neuroscience is no longer driven by loose analogies or metaphors. Its practice reveals that the main aspects of the analogy between the computer and the brain are used for methodological purposes. In particular, Marr's approach to computational explanation, along with a mechanistic stress on the discovery of underlying operations, and the relevance of all neuronal detail for the phenomenon at hand, have made it a successful research tradition.

The unity of computational neuroscience can be seen mostly on the methodological level. Theoretically, computational neuroscientists are divided, defending many competing approaches. They all have their own preferred phenomena of interest and favorite modeling paradigms. They also share principles of computationalism, as rephrased mechanistically in Sect. 3, but their work goes beyond mere principles.

This is also why the loose speculations of the 1940s or 1950s have gradually evolved into biologically constrained modeling techniques that continue to deliver insight into the neurocognitive mechanisms.

5 Possible Objections

In this section, I enumerate several objections against the claim defended here. I claim that computationalism is a progressive research tradition that has led to a remarkable development in computational neuroscience. The conceptual connections between computation and cognition, understood in line with the mechanistic approach to physical computation, provide important research insights within empirical research into the computational modeling of nervous systems. All these claims could be undermined by saying that computational neuroscience is not mechanistic, is empirically vacuous, or is immature.

5.1 Not All Computational Neuroscience is Mechanistic

Some researchers claim that there are examples of computational models in neuroscience that remain at odds with the mechanistic approach to explanation. For example, Chirimuuta (2014) claimed that so-called canonical computations are understood in a heavily idealized and non-mechanistic fashion.

But this paper does not aim to establish that all research in computational neuroscience proceeds in the mechanistic fashion. The claim is, rather, that mechanistic research makes the most sense of theoretical and methodological conundrums faced by modelers and takes the issue of connecting the neural evidence with computational modeling most seriously. Of course, not all modeling is currently up to the task, because empirical evidence is either not there, or remains merely consistent with it. However, the defenders of the mechanistic approach claim that modelers should strive to fulfill mechanistic norms of explanation, even if most research will fail to do so for a long time to come. J. Z. Young repeatedly stressed that crucial detail is missing, and more experimental data is needed. He was well aware that he did not establish, for example, that the distributed connectionist model of recurrent networks sustains working memory in the octopus. At the same time, his functional anatomy has features typical of mechanistic explanations. It describes biological mechanisms responsible for phenomena of his interest, and these mechanisms are explained in terms of their component entities and operations. The missing part in his work on memory is a complete description of the computational machinery, as well as mapping this description onto biological mechanisms, which is a requirement of adequate explanatory models in computational neuroscience (Kaplan 2011). This requirement is fulfilled, to a much larger extent, by Eliasmith's Spaun (cf. Miłkowski 2016b).

Note that the semantic view on physical computation by itself has no stance on whether the models proposed by Young are satisfactory or worse than Eliasmith's. It simply has no resources to deal with this issue.

5.2 Indeterminacy Problems are Still Not Solved

Another objection to computational neuroscience, sometimes voiced by other cognitive scientists, is that modeling is empirically unconstrained and, at least to some extent, arbitrary, because there are too many ways to relate it to evidence. One could even go so far as to claim that anything might make any computational model true, because one could describe anything as a computer (Putnam 1991; Searle 1992).

However, current practice shows that the validation of models is much more constrained, as Sect. 4 above shows. It is simply an unacceptable practice to change the target of the model ad hoc when building a computational simulation. A paper that starts with a task of simulating, say, a single dendrite and ends up only simulating the dynamics of water flowing in a pipe (and not a dendrite), would be rejected in any respectable journal. Moreover, if this model described not only a single dendrite, but also some other phenomena, this coincidence would not be detrimental. The only detrimental situation would be one in which all possible targets fitted the model. In other words, it would be a real disaster if all possible algorithms simultaneously could be truly ascribed to any physical system. But, as many theorists have pointed out (Chalmers 2011; Miłkowski 2013; Piccinini 2015), the arguments to that conclusion are all unsound, because they assume that all there is to physical computation is a simple mapping between a mathematical description of computation, and some physical state of affairs. That issue, however, goes beyond the scope of this paper.

5.3 Computational Neuroscience is Not Mature Enough

In Sect. 4.4, it was admitted that computational neuroscience remains theoretically fragmented. This, one could contend, means that it is not a mature science in the sense of Thomas Kuhn (1962). It is still full of anomalies. So, as a matter of actual fact, computational neuroscience has not progressed much.

This objection assumes that the notion of mature science can be properly used to assess scientific disciplines. But this is not at all clear. Any lively scientific field of research includes a fair amount of controversy, and this is not a sign of degeneracy. Moreover, as Laudan (1977, p. 151) has noticed, Kuhn was unable to point to any major science in which paradigm monopoly has been the rule, or in which the foundational debate was absent. Science has no disdain for anomalies.

6 Conclusion

In this paper, I argued that the computationalism is a research tradition that should not be conflated with one of the component approaches or theories in this tradition. The term *the computational theory of mind* is a misnomer. It is not a single theory, rather a series of metaphysical and methodological assumptions that are shared by particular, sometimes conflicting, computational theories (note the plural). These

assumptions have formed the core of computational cognitive science and early cybernetic efforts to model the brain, which have then transformed slowly into what we now know as computational neuroscience. This can be seen in the development of J. Z. Young's work on memory units in the octopus, which still does not fulfill the methodological standards that are met by contemporary models, such as the model of working memory using SPA that relates behavioral, neuroscientific, and computational data. Even if the short sketch of this transformation remains somewhat a simplified caricature, it retains, or so I would claim, the most important features of this research tradition. For all we can say right now, it has turned out to be fruitful, and will be useful in formulating future theories of the brain and cognition.

The metaphysical assumptions of the tradition can be defended in two ways: by appealing to conceptual analogy arguments, and by showing how these analogy arguments have fostered the development of a progressive research tradition, which has already gone far beyond mere handwaving and speculation. Loose analogies between some brain and some type of computer are no longer deemed to be cognitively satisfactory. Even if analogies have to be understood in broader terms than were presupposed in the 1970s or 1980s, so as not to exclude a priori any of the possible computational theories of brain computation, understanding these assumptions more broadly does not lead to methodological flimsiness. Computational modeling is required to be explanatory, and to provide further insight into phenomena of interest. And it does. It shows, for example, that some yet-unobserved behavioral features are to be expected, which can then be experimentally tested.

Moreover, best practices in computational neuroscience already aim at fulfilling the norms of mechanistic explanation, which requires that all and only causally relevant factors should be included in the models, as well as the complete mapping between computational descriptions and mechanism components. All this vindicates the claim that computationalism is not only here to stay, it becomes stronger every year.

Acknowledgements The work on this paper was funded by a National Science Centre (Poland) research Grant under the decision DEC-2014/14/E/HS1/00803. The author wishes to thank three anonymous reviewers for their insightful comments and criticisms.

References

Adams, F., & Aizawa, K. (2008). *The bounds of cognition*. Malden, MA: Blackwell Pub.

Akagi, M. (2017). Rethinking the problem of cognition. *Synthese*. https://doi.org/10.1007/s11229-017-1383-2.

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review, 85*(4), 249. https://doi.org/10.1037/0033-295X.85.4.249.

Anderson, J. R., Carter, C. S., Fincham, J. M., Qin, Y., Ravizza, S. M., & Rosenberg-Lee, M. (2008). Using fMRI to test models of complex cognition. *Cognitive Science, 32*(8), 1323–1348. https://doi.org/10.1080/03640210802451588.

Apter, M. (1970). *The computer simulation of behaviour*. London: Hutchinson.
Baddeley, A. (1998). Recent developments in working memory. *Current Opinion in Neurobiology, 8*(2), 234–238. https://doi.org/10.1016/S0959-4388(98)80145-1.
Bechtel, W. (2008). Mechanisms in cognitive psychology: What are the operations? *Philosophy of Science, 75*(5), 983–994.
Bhattacharya, B. S., & Chowdhury, F. N. (Eds.). (2015). *Validating neuro-computational models of neurological and psychiatric disorders* (Vol. 14). Cham: Springer. https://doi.org/10.1007/978-3-319-20037-8.
Boden, M. A. (2014). GOFAI. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 89–107). Cambridge: Cambridge University Press.
Boycott, B. B. (1998). John Zachary Young. 18 March 1907–4 July 1997. *Biographical Memoirs of Fellows of the Royal Society, 44,* 487–509. https://doi.org/10.1098/rsbm.1998.0031.
Buckner, C. (2015). A property cluster theory of cognition. *Philosophical Psychology, 28*(3), 307–336. https://doi.org/10.1080/09515089.2013.843274.
Chalmers, D. J. (2011). A computational foundation for the study of cognition. *Journal of Cognitive Science, 12,* 325–359.
Chemero, A. (2003). Information for perception and information processing. *Minds and Machines, 13,* 577–588.
Chirimuuta, M. (2014). Minimal models and canonical neural computations: The distinctness of computational explanation in neuroscience. *Synthese, 191*(2), 127–153. https://doi.org/10.1007/s11229-013-0369-y.
Chomsky, N. (1995). Language and Nature. *Mind, 104*(413), 1–61. https://doi.org/10.1093/mind/104.413.1.
Chrisley, R. (2000). Transparent computationalism. In M. Scheutz (Ed.), *New computationalism: Conceptus-studien 14* (pp. 105–121). Sankt Augustin: Academia Verlag.
Collier, J. D. (1999). Causation is the transfer of information. In H. Sankey (Ed.), *Causation, natural laws and explanation* (pp. 279–331). Dordrecht: Kluwer.
Coltheart, M. (2013). How can functional neuroimaging inform cognitive theories? *Perspectives on Psychological Science, 8*(1), 98–103. https://doi.org/10.1177/1745691612469208.
Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science, 1*(2), 89–97. https://doi.org/10.1080/00207727008920220.
Craik, K. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.
Craver, C. F. (2007). *Explaining the Brain. Mechanisms and the mosaic unity of neuroscience*. Oxford: Oxford University Press.
Cummins, R. (1996). *Representations, targets, and attitudes*. Cambridge, MA: MIT Press.
Deng, C., Yuan, H., & Dai, J. (2017). Behavioral manipulation by optogenetics in the nonhuman primate. *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry*. https://doi.org/10.1177/1073858417728459.
Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy, 68*(4), 87–106.
Dennett, D. C. (1984). Cognitive wheels: The frame problem of AI. In C. Hookway (Ed.), *Minds, machines and evolution* (pp. 129–151). Cambridge: Cambridge University Press.
Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
Dennett, D. C. (1991). Real patterns. *Journal of Philosophy, 88*(1), 27–51.
Dennett, D. C. (1993). Book review: Allen Newell, unified theories of cognition. *Artificial Intelligence, 59,* 285–294.
Downey, A. (2017). Predictive processing and the representation wars: A victory for the eliminativist (via fictionalism). *Synthese*. https://doi.org/10.1007/s11229-017-1442-8.
Drescher, G. L. (1991). *Made-up minds: A constructivist approach to artificial intelligence*. Cambridge, MA: MIT Press.
Dresner, E. (2010). Measurement-theoretic representation and computation-theoretic realization. *The Journal of Philosophy, 107*(6), 275–292.
Dretske, F. I. (1986). Misrepresentation. In R. Bogdan (Ed.), *Belief: Form, content, and function* (pp. 17–37). Oxford: Clarendon Press.
Egan, F. (2010). Computational models: A modest role for content. *Studies In History and Philosophy of Science Part A, 41*(3), 253–259. https://doi.org/10.1016/j.shpsa.2010.07.009.
Eliasmith, C. (2013). *How to build the brain: A neural architecture for biological cognition*. New York: Oxford University Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., et al. (2012). A large-scale model of the functioning brain. *Science, 338*(6111), 1202–1205. https://doi.org/10.1126/science.1225266.

Farah, M. J. (1988). Is visual imagery really visual? Overlooked evidence from neuropsychology. *Psychological Review, 95*(3), 307–317. https://doi.org/10.1037/0033-295X.95.3.307.

Fitch, W. T. (2014). Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition. *Physics of life reviews, 11*(3), 329–364. https://doi.org/10.1016/j.plrev.2014.04.005.

Fodor, J. A. (1968). *Psychological explanation: An introduction to the philosophy of psychology*. New York: Random House.

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese, 28*(2), 97–115. https://doi.org/10.1007/BF00485230.

Fodor, J. A. (1975). *The language of thought* (1st ed.). New York: Thomas Y. Crowell Company.

Fodor, J. A. (1984). Semantics, Wisconsin style. *Synthese, 59*(3), 231–250. https://doi.org/10.1007/BF00869335.

Fresco, N. (2010). Explaining computation without semantics: Keeping it simple. *Minds and Machines, 20*(2), 165–181. https://doi.org/10.1007/s11023-010-9199-6.

Frixione, M. (2001). Tractable competence. *Minds and Machines, 11,* 379–397.

Garzon, F. C. (2008). Towards a general theory of antirepresentationalism. *The British Journal for the Philosophy of Science, 59*(3), 259–292. https://doi.org/10.1093/bjps/axl007.

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: Causally relevant and different from detectors. *Biology and Philosophy, 32,* 337–355. https://doi.org/10.1007/s10539-017-9562-6.

Glennan, S. (2017). *The new mechanical philosophy*. New York, NY: Oxford University Press.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42,* 335–346.

Haugeland, J. (1985). *Artificial intelligence: The very idea*. Cambridge, MA: MIT Press.

Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge, MA: MIT Press.

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese, 183*(3), 339–373. https://doi.org/10.1007/s11229-011-9970-0.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. *Can Theories be Refuted?*. Dordrecht: Springer. https://doi.org/10.1007/978-94-010-1863-0_14.

Laudan, L. (1977). *Progress and its problem: Towards a theory of scientific growth*. Berkeley, CA: University of California Press.

Lloyd, S. (2002). Computational capacity of the universe. *Physical Review Letters, 88*(2), 1–17. https://doi.org/10.1103/PhysRevLett.88.237901.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.

MacKay, D. M. (1969). *Information, mechanism and meaning*. Cambridge: M.I.T. Press.

Maley, C. J. (2017). Toward analog neural computation. *Minds and Machines*. https://doi.org/10.1007/s11023-017-9442-5.

Marr, D. (1982). *Vision. A computational investigation into the human representation and processing of visual information*. New York: W. H. Freeman and Company.

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. Dordrecht: Reidel.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5,* 115–133.

Meyer, D. E., Osman, A. M., Irwin, D. E., & Yantis, S. (1988). Modern mental chronometry. *Biological Psychology, 26*(1–3), 3–67. https://doi.org/10.1016/0301-0511(88)90013-0.

Miłkowski, M. (2011). Beyond formal structure: A mechanistic perspective on computation and implementation. *Journal of Cognitive Science, 12*(4), 359–379.

Miłkowski, M. (2012). Limits of computational explanation of cognition. In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence* (pp. 69–84). Berlin: Springer. http://www.springerlink.com/content/k6w34j70459wv782/.

Miłkowski, M. (2013). *Explaining the computational mind*. Cambridge, MA: MIT Press.

Miłkowski, M. (2016a). Models of Environment. In R. Frantz & L. Marsh (Eds.), *Minds, models and milieux* (pp. 227–238). London: Palgrave Macmillan. http://www.palgrave.com/page/detail/minds-models-and-milieux-/?isb=9781137442499.

Miłkowski, M. (2016b). Explanatory completeness and idealization in large brain simulations: A mechanistic perspective. *Synthese, 193*(5), 1457–1478. https://doi.org/10.1007/s11229-015-0731-3.
Miłkowski, M. (2017a). Objections to Computationalism. A Short Survey. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society. Computational Foundations of Cognition* (pp. 2723–2728). Presented at the 39th Annual Meeting of the Cognitive Science Society, London: Cognitive Science Society. https://mindmodeling.org/cogsci2017/papers/0515/index.html.
Miłkowski, M. (2017b). The false dichotomy between causal realization and semantic computation. *Hybris, 38,* 1–21.
Mole, C., & Klein, C. (2010). Confirmation, refutation, and the evidence of fMRI. In S. J. Hanson & M. Bunzl (Eds.), *Foundational issues in human brain mapping* (pp. 99–112). Cambridge: The MIT Press. https://doi.org/10.7551/mitpress/9780262014021.003.0010.
Neisser, U. (1963). The imitation of man by machine: The view that machines will think as man does reveals misunderstanding of the nature of human thought. *Science, 139*(3551), 193–197. https://doi.org/10.1126/science.139.3551.193.
Nelson, R. (1987). Machine models for cognitive science. *Philosophy of Science, 54*(3), 391–408.
Newell, A. (1980). Physical symbol systems. *Cognitive Science: A Multidisciplinary Journal, 4*(2), 135–183. https://doi.org/10.1207/s15516709cog0402_2.
Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*(3), 113–126. https://doi.org/10.1145/360018.360022.
O'Brien, G., & Opie, J. (1999). A defense of cartesian materialism. *Philosophy and Phenomenological Research, 59*(4), 939–963. https://doi.org/10.2307/2653563.
Osman, M. (2014). *Future-minded: The psychology of agency and control*. Houndmills: Palgrave Macmillan.
Pascual-Leone, A., Walsh, V., & Rothwell, J. (2000). Transcranial magnetic stimulation in cognitive neuroscience—Virtual lesion, chronometry, and functional connectivity. *Current Opinion in Neurobiology, 10*(2), 232–237. https://doi.org/10.1016/S0959-4388(00)00081-7.
Piccinini, G. (2008). Computation without representation. *Philosophical Studies, 137*(2), 205–241. https://doi.org/10.1007/s11098-005-5385-4.
Piccinini, G. (2011). The physical church-turing thesis: Modest or bold? *The British Journal for the Philosophy of Science, 62*(4), 733–769. https://doi.org/10.1093/bjps/axr016.
Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford: Oxford University Press.
Piccinini, G., & Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive Science, 37*(3), 453–488. https://doi.org/10.1111/cogs.12012.
Posner, M. I. (2005). Timing the brain: Mental chronometry as a tool in neuroscience. *PLoS Biology, 3*(2), e51. https://doi.org/10.1371/journal.pbio.0030051.
Putnam, H. (1991). *Representation and Reality*. Cambridge, MA: The MIT Press.
Pylyshyn, Z. W. (1973). The role of competence theories in cognitive psychology. *Journal of Psycholinguistic Research, 2*(1), 21–50. https://doi.org/10.1007/BF01067110.
Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. Cambridge, MA: MIT Press.
Ramsey, W. M. (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.
Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (Vol. 2, pp. 64–99). New York: Appleton-Century-Crofts.
Rooij, I., Wright, C. D., & Wareham, T. (2010). Intractability and the use of heuristics in psychological explanations. *Synthese*. https://doi.org/10.1007/s11229-010-9847-7.
Rowlands, M. (2009). Extended cognition and the mark of the cognitive. *Philosophical Psychology, 22*(1), 1–19. https://doi.org/10.1080/09515080802703620.
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(03), 1–19. https://doi.org/10.1017/S0140525X00005756.
Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.
Shagrir, O. (2010a). Brains as analog-model computers. *Studies In History and Philosophy of Science Part A, 41*(3), 271–279. https://doi.org/10.1016/j.shpsa.2010.07.007.
Shagrir, O. (2010b). Towards a modeling view of computing. In G. Dodig-Crnkovic & M. Burgin (Eds.), *Information and computation*. Singapore: World Scientific Publishing.
Shanahan, M. (1997). *Solving the frame problem: A mathematical investigation of the common sense law of inertia*. Cambridge, MA: MIT Press.

Shapiro, S. (2006). Computability, proof, and open-texture. In A. Olszewski, J. Woleński, & J. Robert (Eds.), *Church's thesis after 70 years* (pp. 420–455). Berlin: Springer.

Shieber, S. M. (1985). Evidence against the context-freeness of natural language. *The formal complexity of natural language* (pp. 320–334). Dordrecht: Springer. https://doi.org/10.1007/978-94-009-3401-6_12.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129–138.

Simon, H. A. (1993). The human mind: The symbolic level. *Proceedings of the American Philosophical Society, 137*(4), 638–647.

Sloman, A. (1978). *The computer revolution in philosophy: Philosophy, science, and models of mind*. Atlantic Highlands N.J.: Humanities Press.

Sprevak, M. (2013). Fictionalism about neural representations. *The Monist, 96*(4), 539–560.

Steels, L. (2008). The symbol grounding problem has been solved, so what' s next? In M. de Vega, A. M. Glenberg, & A. C. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 223–244). Oxford: Oxford University Press.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning, 3*(1), 9–44. https://doi.org/10.1007/BF00115009.

Taylor, W. K. (1964). Cortico-thalamic organization and memory. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 159*(976), 466–478. https://doi.org/10.1098/rspb.1964.0014.

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

Villalobos, M., & Dewhurst, J. (2017). Why post-cognitivism does not (necessarily) entail anti-computationalism. *Adaptive Behavior, 25*(3), 117–128. https://doi.org/10.1177/1059712317710496.

Von Neumann, J. (1958). *The computer and the brain*. New Haven: Yale University Press.

Wheeler, M. (2005). *Reconstructing the cognitive world*. Cambridge, MA: MIT Press.

Wood, W., & Neal, D. T. (2007). A new look at habits and the habit-goal interface. *Psychological Review, 114*(4), 843–863. https://doi.org/10.1037/0033-295X.114.4.843.

Young, J. Z. (1964). *A model of the brain*. Oxford: Clarendon.

Young, J. Z. (1965). The Croonian Lecture, 1965—The organization of a memory system. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 163*(992), 285–320. https://doi.org/10.1098/rspb.1965.0071.

Young, J. Z. (1978). *Programs of the brain*. Oxford: Oxford University Press.

Young, J. Z. (1991). Computation in the Learning System of Cephalopods. *The Biological Bulletin, 180*(2), 200–208. https://doi.org/10.2307/1542389.