Your Brain is Like a Computer:
Function, Analogy, Simplification

M. Chirimuuta (mac289@pitt.edu)
History & Philosophy of Science, University of Pittsburgh

ABSTRACT
The relationship between brain and computer is a perennial theme in theoretical neuroscience, but it has received relatively little attention in the philosophy of neuroscience. This paper argues that much of the popularity of the brain-computer comparison (e.g. circuit models of neurons and brain areas since McCulloch and Pitts [1943]) can be explained by their utility as ways of simplifying the brain. More specifically, by justifying a sharp distinction between aspects of neural anatomy and physiology that serve information-processing, and those that are 'mere metabolic support,' the computational framework provides a means of abstracting away from the complexities of cellular neurobiology, as those details come to be classified as irrelevant to the (computational) functions of the system. I argue that the relation between brain and computer should be understood as one of analogy, and consider the implications of this interpretation for notions of multiple realisation. I suggest some limitations of our understanding of the brain and cognition that may stem from the radical abstraction imposed by the computational framework.

0. PREAMBLE: LEIBNIZ THE INVENTOR

Many histories of computation begin with the unrealised ambition of Gottfried Leibniz to devise a "universal characteristic", a symbolic language in which factual propositions could be represented and further truths inferred by means of a mechanical calculating device (Davis 2000). Amongst the 20th century pioneers of computer science and artificial intelligence who took Leibniz for an inspirational figure[1] were Warren McCulloch and Walter Pitts (Lettvin 2016: xix). Single cell neurophysiology and the engineering of digital computers both grew into maturity in the early 1940's, and significantly influenced one another (Arbib 2016). Cybernetics – the study of information flow and self-regulation in all systems, living and manufactured – was the natural product of these interconnected developments,[2] while McCulloch and Pitts (1943) opus – "A Logical Calculus of the Ideas Immanent in Nervous Activity" – could plausibly be received as the fruit of Leibniz's 270 year old insight that one and the same power of reasoning may inhabit the living man and the mechanical device (Morar 2015:126 fn11).

[1] See Morar (2015) on Leibniz's invention of a mechanical calculator for the four arithmetical functions, and the history of reception of Leibniz's contributions in this area.

[2] See Kline (2015) and Pickering (2010) for overviews of the cybernetic movement in the USA and UK, respectively.

By showing that, under certain assumptions, small assemblies of connected neurons could be taken to operate as logic gates, McCulloch and Pitts were able to claim that the brain *is* – not metaphorically or analogously – a computer. However, the prospect that logic by itself would be all the theory needed to understand the brain turned out to be a mirage. According to the recollections of neurophysiologist Jerome Lettvin, the results of detailed observation of the responses of neurons in the frog's retina left Pitts severely disillusioned because the peculiarities of neuronal behaviour did not make sense from a purely logical point of view.[3]

Following the early literalism, and the subsequent apprehension that the biological system is more tangled than the crystalline ideals of logicians would have it, the relation between brain and computer has been left under-specified. Computer models of neural systems are more than mere models in the sense of simulations, like weather models, that represent but do not re-enact the processes of nature. Instead, neural circuits, and the computational models of them, are thought by the scientists to *be doing the same thing* – processing information (Miłkowski 2018). At the same time, many have voiced the concern that the electronic computer is a mere metaphor for the biological brain, one that places a conceptual box around neuroscientists' thinking and should be discarded along with the hydraulic model of the nervous system, and the image of the cortex as a telephone exchange (Daugman 2001).

In this paper I account for the tenacity of the idea of brain as a computer by appealing to its usefulness as a means of simplifying the brain. I will take the brain-computer relationship to be one of analogy, whereby comparisons are drawn between electronic systems -- engineered to be somewhat functionally similar to biological ones -- and the vastly more complex organic brain. In particular, the brain-computer analogy permits scientists to draw a distinction between the aspects of neuro-anatomy and physiology that are "*for information processing*", as opposed to "*mere metabolic support*". The analogy offers answers to the question of what neural mechanisms are *for,* which are left hanging if one takes the brain only to be an intricate causal web, and one neglects the functional perspective afforded by thinking of the brain as an organic computer. This makes research in neurobiology more efficient by channelling the possibly endless delineation of biochemical interactions along the paths carved out by hypotheses arrived at by reverse engineering the information-processing functions of the neurons.

[3] "up to that time [of results of Lettvin et al. (1959)], Walter had the belief that if you could master logic, and really master it, the world in fact would become more and more transparent. In some sense or another logic was literally the key to understanding the world.

It was apparent to him after we had done the frog's eye that even if logic played a part, it didn't play the important or central part that one would have expected." Lettvin, interviewed in Anderson and Rosenfeld (1998: 10)

1. SIMPLIFICATION AND THE COMPUTATIONAL BRAIN

As stated above, my view is that the relationship between brain and electronic computer, neural physiology and patterns of activation in a circuit board, should be interpreted as one of analogy. This is in contrast with the view that the brain *is literally* a kind of computer, and that neural circuits are one of many potential realisers for the coding schemes discovered by computational neuroscientists, and sometimes implemented by AI engineers (the ones aiming at biological realism). In Section 2 I give a proper elaboration of this contrast, and state some advantages of my own interpretation. The claim of this section is that a major benefit of computational theory in neuroscience is the simplification of the brain that it affords. What I say here is neutral between the literal and analogical interpretations of computational models of the brain (regardless of whether the modellers whose work I discuss themselves understand their models more literally or analogically).

We have noted already that the earliest hopes for a computational theory of the brain – McCulloch and Pitts' plan for neural reverse engineering on the assumption that the brain is a Turing machine and made up of neuronal logic gates (Piccinini 2004) – were defeated by the unruliness (with respect to McCulloch and Pitts' logically derived expectations) of the responses of actual neurons to visual stimulation. Given these initial disappointments, one might ask how it was that computationalism still went on to become the dominant theoretical framework for neuroscience.[4] This is a broad question which deserves a complex answer, referring to historical and sociological factors, and to differences between sub-specialities within the science. However, for the purposes of this paper, I offer a simple answer, that boils down only to one characteristic of computationalism – that it provides neuroscientists with a very useful, possibly indispensable, means to simplify their subject of investigation. More specifically, my claims are (1) that computationalism permits a distinction between the functional (information processing) aspects of neural anatomy and physiology and what is there merely as metabolic support, thereby justifying the neglect of countless layers of biological complexity; and (2) that computational theory, in giving the specification of neural functions, provides an ingredient lacking in purely mechanistic approaches to neurobiology, without which it would be far more difficult to separate relevant from irrelevant causal factors and hence to state when the characterisation of a mechanism is sufficiently complete.

[4] Note that this is should not be confused with the issue of whether the dominant mode of explanation in neuroscience is mechanistic or computational. Those on the mechanist side of this debate, such as Kaplan (2011), acknowledge the importance of computationalism in theoretical neuroscience, and argue furthermore that computational models provide mechanistic explanations. Another point is that those promoting dynamical systems theory as a better theoretical framework than computationalism for some neural systems (e.g. Shenoy, Sahani, and Churchland (2013)) do not dispute the dominance of computationalism in neuroscience as it stands.

1.1 The Isolation of the Functional

It should not be news to anyone who has observed the practice of science that part of the task (and art) of devising a new experiment or explanation is the drawing of a distinction between the target of investigation and the additional factors that can reasonably be classified as background conditions. For a system of any complexity (which is all of the systems studied in biological science), the outcome of the endeavour largely turns on the aptness of the distinction. As the neurologist Kurt Goldstein (1938) argued, all of the supposed "background" factors within an organism are highly relevant to the behaviour of the whole creature, in ways that most of experimental biology ignores; yet even if one acknowledges the lack of an absolute distinction between target and background, it is still usually appropriate for the biologist to train her attention selectively on the target, as one does with a visual image affording figure-ground separation.

My contention here is that much of the value that the computational framework provides to neuroscience is in the distinction it supports between the function of a neural system (information processing), which provides the target of investigation, and the residual features that can be placed in the background as mere metabolic support. The classic characterisation of the neuron as a device which gathers inputs at the dendrites, calculates a function and delivers an output (a number of spikes sent down the axon) is the most prevalent way that this distinction has been put to use in neuroscience. While this picture is much broader than McCulloch and Pitts' (1943) formalism, they can be credited with disseminating the idea that the single neuron is an input-output device, and giving neuro-modellers an excuse for abstracting away from most of the cell biology underling the reception and generation of action potentials:

> The liberating effect of the mode of thinking characteristic of the McCulloch and Pitts theory can be felt on two levels. ….. On the local level it eliminates all consideration of the detailed biology of the individual cells from the problem of understanding the integrative behaviour of the nervous system. This is done by postulating a hypothetical species of neuron defined entirely by the computation of an output as a logical function of a restricted set of input neurons. (Papert 2016: xxxiii)

The utility of this simple picture goes a long way to explaining the persistence of the "neuron doctrine"—the thesis that neurons are the functional unit of the nervous system, whose job it is to receive, process and send information—in the face of some countervailing empirical findings (Bullock et al. 2005).[5]

[5] Cao (2014) recommends going beyond the neuron doctrine to consider synapses and glia also as functional units of the nervous system. This raises the question of the technical feasibility of gathering synapse-resolution data of neural responses, and attempting to model the brain in such a fine-grained way (noting that each cortical neuron receives, on average, tens of thousands of inputs). If the neuron doctrine provides a "good enough" framework for modelling the brain, especially useful for the activation patterns associated with observable behaviours (perception, learning, decision making) which involve large populations of neurons, then there is little reason to

The strategy, just outlined, for isolating the functional begins with the concrete neural system and abstracts away from it all features classified as non-functional, metabolic support. Another modus operandi is to start with the specification of a cognitive task (such as detection of edges in a photograph), consider what computations would be needed to achieve the task, and then to build an artificial system (i.e. a computational model) that performs it. With the model in place, the final step is to use it as a template or map when looking for activation and connectivity patterns in the brain that are responsible for the performance of this task. This strategy is described by Lettvin, in response to the criticism that computational models used in neuroscience – such as connectionist networks – lack similarity to neural systems:

> But, even if ideally one could record from any element or part of an element in situ, it is not in the least obvious how the records could be interpreted.[6] To a greater degree than in any other current science, we must know what to look for in order to recognize it…..
>
> This is where a prior art is needed, some understanding of process[7] design. And that is where AI, PDP, and the whole investment in building [neurocomputational models of intelligence] enter in. Critics carp that the current golems do not resemble our friends Tom, Dick, or Harry. But the brute point is that a working golem is not only preferable to total ignorance, it also shows how processes can be designed analogous to those we are frustrated in explaining in terms of nervous action. It also suggests what to look for. Lettvin (2016:xvii- xviii)[8]

If anything, the problem of "knowing what to look for" is more acute now than when Lettvin wrote this. In the last ten years, the increase in the variety of tools and methods for observing neural activity (from single cells to whole brains) has surprised and delighted many. However, the downside of these advances is that they bring to light kinds of complexity that were not previously apparent, especially at sub-cellular scales. This is how neuroscientist Yves Frégnac describes the situation:

attempt the impossible and replace neurons with synapses as the fundamental signalling systems, even if one acknowledges that in the brain much information processing does occur within synapses. Below I take up the issue of the importance of these details that are relegated to the background in the classic neuro-computational picture.

[6] A point made vivid by Jonas and Kording (2017)

[7] Lettvin often uses this word in his characterisation of the 'engineering-stance' in neuroscience. It should not be confused with the notion of "process models" in psychology, or other kinds of mechanistic models.

[8] Pickering (2011:6) takes this methodology to be the standard practice for cybernetics in neuroscience, though many of the artificial devices where not computer programmes.

> Just how did the cyberneticians attack the adaptive brain? The answer is, in the first instance, by building electromechanical devices that were themselves adaptive and which could thus be understood as perspicuous and suggestive models for understanding the brain itself. The simplest such model was the servomechanism—an engineering device that reacts to fluctuations in its environment in such a way as to cancel them out. A domestic thermostat is a servomechanism; so was the nineteenth-century steam-engine 'governor' which led Wiener to the word 'cybernetics.'

> Each overcoming of technological barriers opens a Pandora's box by revealing hidden variables, mechanisms, and nonlinearities, adding new levels of complexity. By reaching the microscopic-scale resolution, advanced technologies have unveiled a new world of diversity and randomness, which was not apparent in pioneer functional studies using spike rate readout or mesoscopic imaging of reduced sensitivity. ((Frégnac)2017:471)

He points to the need for a greater understanding of how mesoscopic and macroscopic regularities emerge from the processes observed microscopically. But a wider point is that if artificial systems, sharing none of the microscopic details of the neural ones, can be built to replicate some specific functions,[9] then one has an acceptable excuse for keeping shut the Pandora's box of sub-cellular neurobiology.

1.2 Mechanism and Function

In response to a criticism of the mechanistic account of explanation, which takes issue with the favouring of more detailed descriptions of mechanisms as providing better explanations than less detailed, 'sketchy' ones, Craver and Kaplan (2018) emphasise that their account has never favoured more detailed descriptions, per se, but has only suggested that models describing more of the *relevant* details may have the edge over more abstract ones. But this immediately raises the question of how the scientist comes to know how to distinguish the relevant from the irrelevant factors. In any biological system, the nervous system especially, one finds a densely inter-connected causal web with many layers of structural intricacy, and patterns of effect across various spatial and temporal scales. Craver and Kaplan appeal to a "mutual manipulability" criterion that is clear and unobjectionable in principle.[10] However, in practice it is hard to see how only the causal factors in a neural system relevant to a particular phenomenon will be isolated if only the mechanistic strategy is employed. An individual neuron will have thousands of feasible targets or 'handles' for experimental manipulation – for example, the different kinds of ion channels, which could be blocked on select portions of the membrane; the various different receptors that could be agonised or antagonised; the countless proteins transcribed in the cell which could be targets of genetic manipulation. One needs to multiply this list of causal variables by 10 or by 100 if the mechanism comprises a small population of neurons. One faces a combinatorial explosion of experiments that would be

[9] I am alluding here to multiple realisation – a topic to be discussed directly in Section 2. But the point can still be made without supposing there are cases in which one would want to say that an artificial and a neural system are two different realisers of *the same* function. Consider just the comparison between a fairly abstract and a highly detailed model of a neural circuit (e.g. a model where neurons are just represented as a time series of spike rates, and a "compartment model" which represents some of the anatomical structure of the neuron). If the former is an equally good working model of the function of interest, then it is a reasonable working assumption that the behaviour of the neural system can be understood without reference to sub-cellular structure.

[10] "A factor is constitutively relevant when (ideal) interventions on putative component parts can be used to change the explanandum phenomenon as a whole and, conversely, interventions on the explanandum phenomenon as a whole can produce changes in the component parts" Craver and Kaplan (2018: 20)

needed to determine the independent causal relevance of each of these factors in a putative mechanism. But of course neuroscientists do not plan sequences of experiments according to brute force search! When designing an experiment to determine which of the many causal variables present in a mechanism are relevant to an explanandum phenomenon, how does a neuroscientist know which ones to select from an inexhaustible list? One should think of hypotheses regarding the information processing functions of neuronal structures as heuristics that drastically reduce this search space.

For example, at a fairly high level of abstraction, only net excitation minus inhibition is the causal factor relevant to determining whether a neuron's firing rate will increase or decrease. This abstraction disregards the kinds of neurotransmitters found at the synapse, receptor types, and location of synapses.[11] And of course this is the kind of abstraction fostered by the neuron doctrine and fundamental to McCulloch and Pitts' vision of the brain as a computer in which the logic gates are built from neurons.[12] In essence, without any prior assumption in place about what the neuron's function is, and what aspects of physiology and anatomy are relevant to it, the search for relevant causal factors would have to proceed by brute force or be guided by pure prejudice. This indicates that the functional, informational processing perspective on neural systems is an indispensable complement to the mechanistic approach in neurobiology.

The difference between the physicist's and the engineer's perspectives on nature is a useful analogue to the difference between mechanistic and computational perspectives in neuroscience (Fairhall 2014). When one considers the structures of the brain as a physical system, it is a web of causal interactions in which considerations of function are alien; in contrast, the notions of design and function are inherent to the engineering perspective, from which it is natural to regard the brain as a target of reverse-engineering (Sterling and Laughlin 2016). The mechanistic approach is supposed only to decompose a system into its structures and causal interactions, showing how their interaction brings about or constitutes the phenomenon which identifies the mechanism. On the computational approach, one begins with the consideration of what the neural system is *for,* and the question of how that function is achieved is addressed only after this. When dealing with complex, biological systems, any

[11] Craver and Kaplan (2018:p.19 fn 16) appeal to the purely causal notion of "screening off" in order to address the question of why complete (ontic) explanations do not end in quarks. The idea is that "low-level differences" will be ignored if they "make no relevant difference once the higher-level behaviour is fixed." I would like to point out that for the kind of abstractions I mention here, screening off should not be expected to occur – i.e. these excluded details do causally affect neuronal behavior in ways that are not fully summarized by the "higher level" variables of net excitation and inhibition, because of non-linearities in the behaviour of the cell. This suggests that a search for "relevant details" that proceeded only by the method of searching for "higher level" causal variables to replace "lower level" ones would not result in the abstractions found to be most useful in computational neuroscience.

[12] There is latitude here in the abstracting assumptions. I have described a case where total inhibition is subtracted from total excitation, whereas McCulloch and Pitts (1943:118) posit that inhibitory input at any one synapse will cancel out the effects of excitation.

attempt to employ only the neutral (function-less) physical stance would quickly get one lost amongst tangled causal details. This is a point made by the neurologist Francis Walshe:

> The modern student finds it difficult to see the wood for the trees … He does not always have a synoptic concept of the nervous system in his mind … If we subject a clock to minute analysis by the methods of physics and chemistry, we shall learn a great deal about its constituents, but we shall not discover its operational principles, that is, what makes these constituents function as a clock. Physics and chemistry are not competent to answer questions of this order, which are an engineer's task … Both modes have their place and limitations; and they complement one another. (Walshe 1961: 131)[13]

It is the task of theory in science to provide the "synoptic concept" of a subject matter, and in neuroscience the computational theory is best developed, though I do not claim that this is the only possible theory of the nervous system.

A wrinkle in the comparison I have drawn between the physicist's approach and mechanistic perspective in biology is that a mechanistic investigation *does* incorporate a notion of function or purpose, that is completely alien to physics. This is because without such a notion one actually cannot delineate a mechanism – mechanisms are mechanisms *for* the phenomena they produce or constitute (Craver and Kaplan 2018:23 fn19). The tension within the mechanistic outlook is that this notion of function has an ambiguous status.[14] On the one hand, purpose or function cannot be thought of as an *inherent* feature of the mechanism in question (which is, officially, just a purposeless causal web of processes which take place according to the laws of physics and chemistry); on the other hand mechanisms are thought of as defined by the things that they do, which is normally understood as the purpose served in the context of the tissue, organ, or organism. This difference is papered over with the thought that one can gesture at Darwinian adaptation and the notion of selected functions to bridge this gap -- even if, in reality, no-one ever attempts to show that every system classified as a mechanism has actually been a target of natural selection, and so has a "proper function". And in fact Craver and Darden (2013: 53-54) deny that the phenomena which identify mechanisms need be proper functions. Thus, the notion of function is an implicit precondition of the mechanistic perspective in biology; but like an embarrassing relative, it is only rarely mentioned.

[13] See also Knuuttila and Loettgers (2014: 79) on the contrast between physics and engineering based approaches within synthetic biology research.

[14] See Canguilhem (2008) for many remarkable observations on the tensions within the mechanistic perspective, regarding the status of function and finality. The problematic idea that there is an exclusive rather than complementary relationship between mechanistic and teleological perspectives in biology, is evident in Craver and Tabery (2017) description of mechanism as a self-contained "scientific worldview":

> Some have held that natural phenomena should be understood teleologically. Others have been convinced that understanding the natural world is nothing more than being able to predict its behavior. Commitment to mechanism as a framework concept is commitment to something distinct from and, for many, exclusive of, these alternative conceptions. If this appears trivial, rather than a central achievement in the history of science, it is because the mechanistic perspective now so thoroughly dominates our scientific worldview.

In relation to this, Jerome Lettvin makes the very interesting point that the engineering perspective is prominent in biology precisely where there is a vacuum left following biologists' attempt to adhere strictly to physical-chemical (and hence purpose-less) perspectives when conceptualising their subject matter:

> Ever since biology became a science at the hands of biochemists it has carefully avoided or renounced the concept of purpose as having any role in the systems observed…. Only the observer may have purpose, but nothing observed is to be explained by it. This materialist article of faith has forced any study of process out of science and into the hands of engineers to whom purpose and process are the fundamental concepts in designing and understanding and optimizing machines. (1998:13)

Lettvin goes on to say that, "we had better use the process [i.e. functional characterisation] to tell what to look for in the mechanism rather than the other way round." (1998:17).

With this in mind, we can appreciate that cybernetics, the scientific movement in which McCulloch and Pitts were players, and from which today's computational neuroscience descended, was self-consciously a science of finality in a mechanistic world. And it was possible for cybernetics to develop as a science of finality because engineering was very well represented in this interdisciplinary research field. Cyberneticians took the design stance in biology, both in the hope of gaining scientific insights, and in order to receive inspiration for the design of intelligent artificial devices. Thus Rosenblueth, Wiener, and Bigelow (1943: 23) simply redefine "teleology" as "purpose controlled by feed-back", and thereby avoid any problematic reference to final causation.[15]

[15] It is worth quoting Rosenblueth, Wiener and Bigelow (1943:23) at length:

> Teleology has been interpreted in the past to imply purpose and the vague concept of a "final cause" has been often added. This concept of final causes has led to the opposition of teleology to determinism. A discussion of causality, determinism and final causes is beyond the scope of this essay. It may be pointed out, however, that purposefulness, as defined here, is quite independent of causality, initial or final. Teleology has been discredited chiefly because it was defined to imply a cause subsequent in time to a given effect. When this aspect of teleology was dismissed, however, the associated recognition of the importance of purpose was also unfortunately discarded. Since we consider purposefulness a concept necessary for the understanding of certain modes of behavior we suggest that a teleological study is useful if it avoids problems of causality and concerns itself merely with an investigation of purpose.

Note also that Francis Walshe, quoted above on the complementary relationship between the physicist's and engineer's stances in neuroscience, was quite critical of Rosenblueth et al.'s paper, highlighting the mismatch between the operation of feedback in the cerebellum and in the artificial system, which, he argues, means the literal interpretation of the cybernetic model is not warranted (Walshe 1951). See also the discussion of Rosenblueth et al. in Mayr (1988).

2. TWO INTERPRETATIONS OF THE BRAIN-COMPUTER RELATIONSHIP

The building of machines in order to elucidate processes underlying vital functions, including cognition, is strategy that goes back at least to the automaton-makers of the eighteenth century.[16] But an open question here is whether, in order to understand the efficacy of this pattern of investigation, one must resort to a literal interpretation of the artificial models (computer programs or other devices) as duplicating and thereby bringing to light the *same* process or function as it occurs in the living system, or if one can still make sense of the research strategy by taking the machine-organism relationship as one of analogy. That is, by saying that the organism is *like* the machine in some determined way, but making salient the numerous differences (disanalogies) that limit the appropriateness of the machine-organism comparison to the narrow domain of the phenomena explicitly modelled.

Theoretical neuroscience has benefitted from a strategic vagueness on this point – the difficult question of whether the differences between brains and computers are significant disanalogies which restrict the scope of the comparison of the two kinds of system has been deferred indefinitely. According to Lettvin, McCulloch was under no illusion that neural assemblies share all the properties and behaviours of digital logic gates. However, the comparison was appropriate because, Lettvin (2016: xviii-xix) asserts, "there are properties of such connected systems that are more or less independent of the intrinsic nature of the nonlinear elements used, whether gates or neurons". The latitude in the "more or less independent" here is useful for the scientist because the observation of relative independence provides clues to the scientist about which causal factors do not need to be made the target of an experiment, and which details may safely be left out without foreclosing on the possibility that the independence may turn out to fail in some circumstances, and that those neglected details might later be the subject of experiment and modelling. However, philosophers of mind not satisfied with such vague assertions have built theories in which the independence ("autonomy") of computational descriptions has been treated as a categorical fact, meaning that there is no important disanalogy between information processing as it occurs in electronic and neural tissue. The higher level, functional properties associated with information processing are said to be multiply realised in neurons and logic gates. I will now provide some exposition of this literal way of interpreting computational models of the brain, before offering an alternative that centres on the notion of analogy.

[16] As Canguilhem (1963: 510) describes, "texts, taken from Quesnay, Vaucanson and Le Cat, do not indeed leave any doubt that their common plan was to use the resources of automatism as a dodge, or as a trick with theoretical intent, in order to elucidate the mechanism of physiological functions by the reduction of the unknown to the known, and by complete reproduction of analogous effects in an experimentally intelligible manner."

2.1 The Literal Interpretation: Formal Realism

One point that can be derived from the above discussion of the relationship between the physical and engineering approaches, and the mechanistic and computational perspectives that go with them (Section 1.2), is that the engineering approach in contemporary biology is a distant echo of the Aristotelian tenet that living systems cannot be understood without a first regard to their purposes and their *forms* (patterns of organisation). These notions of form and finality were, according to popular history, banished from science in the 17th century and then, after a long wandering in exile, put mercifully to death by Darwin. Yet, as various philosophers and historians of biology have argued, these ideas are ever present in modern biology, even if going by different names (Allen, Bekoff, and Lauder 1998). I argued above that cybernetics can be understood as a kind of neo-Aristotelian research programme, in that it restores a place for finality in the science of living systems. Some advocates of functionalism in the philosophy of mind have emphasised the Aristotelian aspects of the theory (Nussbaum and Putnam 1992). Although this connection can sometimes be overstretched (Burnyeat 1992), I give the name *formal realism* to the literal stance towards neuro-computational models, which itself can be thought of as a tenet of functionalism, in order to draw attention to its resonance with Aristotle.[17]

In Aristotle's hylomorphism – as applied to living beings – the explanation of how the body is able to do what it does (achieve its ends) is put in terms of the presence of a form inherent in the matter, which together comprise the body. Forms can be thought of, generally, as patterns or principles of organisation, so that when one takes the literal interpretation of computational models of the brain as a modern version of hylomorphism, the relevant forms are computational functions,[18] not "souls" or "animae", and the neural realiser is the matter made intelligent by the presence of the form. Thus the modern formal realist takes computation to be the essence or principle responsible for cognition and underlying intelligent behaviour. So even though the neuroscientists who work in the computational tradition and offer literal interpretations of their

[17] Another tenet of functionalism is the classic account of multiple-realisation which gives the abstract computational "level" of neuro-modelling a robust ontological interpretation. Elsewhere I call this approach MR 1.0 and argue that it be replaced with an ontologically modest view, MR 2.0, which treats the computational as a level of explanation rather than a level of being (Chirimuuta 2018b). MR 2.0 is consistent with the analogical interpretation of computational models offered below (Section 2.2); indeed, the analogical interpretation is intended to be an elaboration of some of the ideas presented in my earlier paper.

[18] We might also consider here the bivalence of the word "function", which has both a mathematical and a biological sense (Longuenesse 2005: 93). Interestingly, the two meanings coincide in formal realism, where the function is at once the mathematical operation computed by the neurons, and the biological purpose of this activity. Note that because the relevant forms in computational neuroscience are mathematical ones, formal realism here has a Platonic as well as an Aristotelian feel: the underlying order of the brain is a mathematical one. Elsewhere I say more about the Platonic dimension (Chirimuuta forthcoming).

models, and any philosophers following in attendance,[19] would not embrace any characterisation of them as adherents to an Aristotelian metaphysics, to the extent that that their research treats computation as the essence of cognition and intelligence, the label of formal realism is apt.[20]

Hylomorphism does not entail multiple realizability – the notion that the one and the same form can inhere in radically different kinds. However, when the relevant forms are mathematical functions, multiple realizability is inevitable because of the fact that the same computation (e.g. multiplication of 653x10) can be performed by a variety of physical realisers, including an artificial computers (mechanical or electronic) or biological tissue. The picture of an abstract mathematical form, finding itself realised in an array of material substrates – breathing intelligence into them, one might say – has had long appeal. According to Morar (2015:126) this is what occurred to Leibniz after his encounter with the famous adding-subtracting machine invented by Pascal:

> As Leibniz came out through the door of Louis XIV's library after seeing the Pascaline, he left behind all of his previous ideas of what a new type of calculator could look like, but not his goals. He had begun thinking about building a machine since at least 1670, two years before he came to Paris, and the challenge was clear: if mortal man had the power to transpose in 'yellow brass' the faculty of mathematical reasoning, there could be no doubt that God had been able to house a 'more general spirit' into the body of animals, giving them life.

While I do not suppose that any defender of formal realism in computational neuroscience owes us an elaborate metaphysics of an Aristotelian or Leibnizian sort, I will say that the view does bring up some challenging metaphysical questions, as well as empirical ones. The view seems to presuppose a realism about mathematical form which is normally associated with a Platonism -- where mathematical abstracta exist outside space and time. At the same time, mathematical operations are taken to be realized in the material brain, which is located in time and space. Are we to think these mathematical forms as inhering in material objects, in the way that Aristotle's notion of form brought Platonic ideas down to earth? The standard answer to this question is to point to the concept of implementation. The pressing challenge, then, is to give an account of the implementation of computational functions in concrete material that

[19] I am not claiming that *all* computational neuroscientists subscribe to the literal interpretation, but I do think that it is the dominant strand of thought within the discipline. Note that complaints that the brain is *not* a computer usually mean that the brain is not a digital, serial machine. Marcus (2015: 209) nicely expresses the dominant view:

> "it is obvious that brains (especially those of vertebrates) *are* computers, in the sense of being systems that operate over inputs and manipulate information systematically. Brains might not be (purely) *digital* computers, their memories may operate under different principles, and they may perform different sorts of operations on the information they encode, but they surely encode information…. Computers are, in a nutshell, systematic architectures that take inputs, encode and manipulate information, and transform their inputs into outputs. Brains are, so far as we can tell, exactly that."

[20] For example I classify Egan (2017), Shagrir (2010) and Shagrir (2018) as formal realists.

does not imply pancomputationalism (Putnam 1988), while showing how the computational level of explanation is autonomous from the implementational one (Ritchie and Piccinini 2018).

Another issue, noted above, is that the view implies the multiple realisability of computations underlying intelligence, and hence multiple realisation as an empirical fact. Polger and Shapiro (2016) present a thorough case that the evidence for multiple realisation is lacking, contrary to the expectations of functionalist philosophers of mind. Of course others have a different opinion, and it is not obvious that the metaphysical challenges are insurmountable (Aizawa 2018). So I am not claiming that the formal realism is untenable just because of these difficulties. However, the fact that these challenges exist provides motivation for development of an alternative, less metaphysically committal interpretation of computational models of the brain.

2.2 The Analogical Interpretation: Formal Idealism[21]

According to Cassirer, the felt need for an explanation of the applicability of mathematics in empirical science that did not depend on any dogmatic metaphysical assertions was Kant's first step along the road to his critical philosophy (Seidengart 2012: 141). To advance towards an alternative to the literal interpretation of computational models in neuroscience, I suggest that we re-tread this path. While the formal realist takes for granted the brute existence of mathematical forms, which are realised equivalently in brains or computers, the *formal idealist*[22] takes the mathematical forms represented in computational models of the brain not to be straightforward discoveries regarding mathematical structure or information processing in the brain, but constructs developed through an arduous process of experimentation, model building, and analogical reasoning. This Kantian proposal is that the mathematical structures which make the brain intelligible us, as an organ whose function is to process information, are to some extent imposed by us onto the neural system and should not be taken as straightforward discoveries of mathematical forms inherent in the system.[23] Since, by hypothesis, our neuro-computational models are not discoveries of the inherent computational capacities of the brain, but are as abstract and idealised as any other models in science, an analogical interpretation of these models is more appropriate than a literal one.[24]

[21] The analogical interpretation should be understood in the specific sense described here, not to be confused with the "analog-model" account of the brain (Shagrir 2010), which I classify as a formal realism.

[22] Kant (1929: B519, note a) gives "formal idealism" as a gloss for "transcendental idealism". The former term draws attention to the point that the idealism in Kant's philosophy is restricted to the way that our knowledge of nature is *formed* or *structured* by our cognitive capacities rather than a structure pre-given in things-in-themselves. Boyle (manuscript) is a very interesting discussion of hylomorphism, without formal realism, in Kant's philosophy.

[23] See also Chirimuuta (forthcoming) for an argument against formal realism, based on the existence of empirically adequate but incompatible mathematical models of certain brain areas.

[24] Elsewhere I present a detailed case study of linear computational models of V1 and M1, showing that the idealizations present in the models are indispensable if the models are to provide us with understanding of the

In the classic account, Hesse (1966) charts the structure of analogical reasoning in science using diagrams which compare two systems (the analogue source and target) along vertical and horizontal axes. For example, the analogical inference that Mars, because of its similarities with the Earth, *may* support life is depicted in Figure 1.

Earth (Source)		**Mars (Target)**
	Known Similarities	
Orbits the sun Has a moon Revolves on axis Subject to gravity		Orbits the sun Has moons Revolves on axis Subject to gravity
	Inferred Similarity	
Supports life	==>	*May* support life

Figure 1: A schematic for analogical reasoning, after Bartha (2016)

Figure 2 offers an example, based on research published by Mante et al. (2013) on perceptual decision making in the prefrontal cortex.[25] The researchers gathered both neurophysiological and behavioural data from monkeys performing a task in which stimuli varied either in colour or in direction of motion, and depending on a contextual cue the monkey had to report on either one of these stimulus dimensions. They also trained a recurrent neural network (RNN) model to perform a virtual equivalent of the experimental task. Through reverse engineering of the trained RNN, the researchers formulated an explanation of how the network was able to accomplish this kind of decision making, turning on the fact that there is a *line attractor* in the low dimensional state space of the network which allows for integration of context dependent information. The researchers observed a number of similarities between the trained RNN and the prefrontal cortex (see Figure 2). On the basis of this it is possible to make the analogical inference that the process underlying the context-dependent perceptual decision, discovered by reverse engineering the RNN, may also occur within the cortex.

This inference is put forward not as conclusive proof, but as a plausible explanation of the biological function that also serves as a hypothesis for future experimental testing. Because of the forward looking aspect of this kind of analogical reasoning, I call it *prospective*. It should be noted that the authors of this research present the RNN as a literal representation of the coding that occurs in the prefrontal cortex, such that the reverse engineering that leads to the discovery of how the task is performed in the model is thereby a discovery of the biological

brain areas. I therefore concur with Potochnik (2017) and Elgin (2004) that the understanding provided by idealized models is non-factive (Chirimuuta, manuscript).

[25] For a more lengthy discussion of this research and the explanations it affords see Chirimuuta (2018a).

process. In contrast, the analogical interpretation is more tentative than this, being sensitive to the open possibility that future discoveries of dissimilarities between brain and model will call into question the validity of the analogical inference.

Computer (Source) *RNN model*		**Brain (Target)** *Prefrontal Cortex*
	Observed Similarities	
Makes context-dependent perceptual decision. Irrelevant sensory information is represented in the population. In the 3D state space, the angle of the 'choice' axis is fixed in relation to the 'colour' and 'motion' axes.		Makes context-dependent perceptual decision. Irrelevant sensory information is represented in the population. In the 3D state space, the angle of the 'choice' axis is fixed in relation to the 'colour' and 'motion' axes.
	Inferred Similarity	
There is a line attractor in the state space, which explains integration of information.	==>	*May* be that there is a line attractor in the state space, which explains integration of information.

Figure 2. *Prospective* pattern of analogical reasoning.

Figure 3 presents a more elaborate kind of analogical reasoning in neuroscience, that I call *abstractive*. This example is taken from David Marr and Shimon Ullman, whose approach to computational modelling in neuroscience has been highly influential.[26] Because of the "behavioural" similarity observed across the systems (the ability to detect edges), and the similarities in patterns of activation in response to edges, the analogical inference is made that neurons in the cat's early visual system – *retinal ganglion cells* (RGC) and neurons in *lateral geniculate nucleus* (LGN) – can be modelled as computing a *Laplacian of Gaussian* function.[27]

[26] Marr and Ullman (1981); Marr (1982: 54-65).

[27] Marr (1982:64) makes the stronger (but hedged) claim that these neurons *are* computing the function:

> "it is not too unreasonable to propose that the $\nabla^2 G$ function is what is carried by the X cells of the retina and lateral geniculate body, positive values being carried by the on-center X cells, and negative values by the off-center X cells."

This amounts to a formal realism, so I do not propose my weaker interpretation of the case as one proposed by Marr himself (see Egan (2017) and Shagrir (2010) for discussions of this example which instead endorse the literal interpretation). That said, I do think Marr can be read as making the abstractive inference. A short biographical note: I first heard of this example during an undergraduate lecture by the late and much missed Tom Troscianko.

Computer (Source) *Laplacian of Gaussian Model*		**Brain (Target)** *LGN or RGC neurons in cat*
	Observed Similarities	
Detects edges in a photo. Characteristic peaks of model output for onset and offset of edges.		Responds to moving edges. Average increases in neural activity for onset and offset of edges.
	Observed Dissimilarities	
Peaks for onset and offset are *symmetrical*. Implemented in digital computer.		Peaks for onset and offset are *asymmetrical*. [Ignored] Is an electrically excitable cell.
	Inferred Similarity	
Model computes Laplacian of Gaussian function.	==>	RGC and LGN neurons can be modelled as computing Laplacian of Gaussian function.
	Abstractive Inference	
==>	*Differences in implementation are not relevant to the particular capacity here investigated.*	

Figure 3: *Abstractive* pattern of analogical reasoning.

In addition to the observation of similar overall behaviour, the dissimilarities in the material substrates of the systems may also be noted and the *abstractive inference* made that these dissimilarities are *not relevant* to the scientist's investigation of the capacity for edge detection.[28] The possibility of this kind of abstraction is a precondition for Marr's (1982:25) distinction between the levels of computational theory and algorithm, and that of implementation. This kind of abstractive inference fits with my account of how it is that computational models aid neuroscientists in the simplification of the brain – the abstractions discussed above can be licensed by this sort of analogy. But by putting this account of abstraction and simplification in the context of a non-literal, analogical approach to interpretation of neuro-computational models, there is no commitment made here to "computational essentialism" about the brain, or to the idea that all the information processing that occurs in the brain *must* be multiply realisable.

Intrigued by the idea that the retina does calculus, I decided to do my final year research project with Tom, and then went on to do graduate research with one of his collaborators. I am still wondering.

[28] NB – the inference is not that the differences in implementation is irrelevant *tout court*, but that they can reasonably be ignored for this kind of investigation of this particular capacity.

One concern here might be that there is no substantial difference between the literal and analogical interpretations: *if the analogical inferences are warranted by findings of similarities between neural and artificial systems, then this means that the two systems have corresponding structures that are (close to) isomorphic to one another*. The point of the literal interpretation is just to say that the similarity is close enough that we can talk of *the same* structure (e.g. function computed) being instantiated in the two systems. Thus confusion arises if one defines analogies as isomorphisms.[29] That is to claim that if some model is an analogue to a target, then there is some structure in it that is isomorphic to a corresponding *inherent* structure in the target. But this is not how I am conceiving of analogies, since on that conception there would be no daylight between the literal and analogical interpretations of neurocomputational models. Rather, on my conception, to say that a model should be interpreted analogically is to say that the target is *like* the model is some way that may turn out to be dependent on the interests of the scientists, and the techniques they employ. The crucial point is that the structure in the target found to be to be relevantly similar to the model is not assumed to be an inherent, human-independent fact about the target.

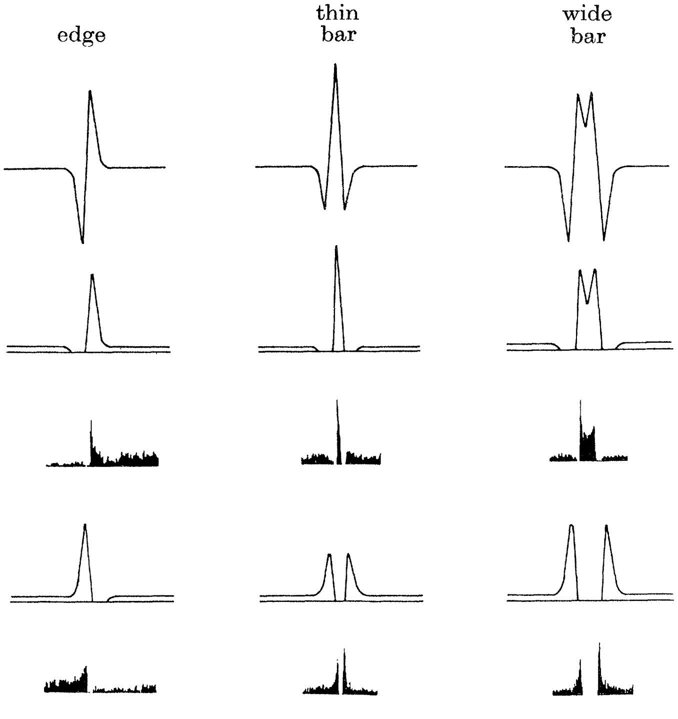

Figure 4 --- need to point out the asymmetries, and how they relate to stimuli.

[29] This is not how analogies are defined in the philosophy of science literature on analogical reasoning. As Dardashti, Thébault, and Winsberg (2017) put it, instances where an isomorphism obtains are a subset of all the cases of analogies in science, and they support stronger inferences than the other cases. See Knuuttila and Loettgers (2014: 87) for further discussion of why analogical reasoning in science goes beyond the isolation of structures that map from model to target.

The distinction between formal realism and formal idealism also provides a useful way to separate the literal and analogical interpretations. According to formal idealism, the relevant similarities are not simply there, waiting to be discovered by the scientist but are in some respect constructed, or massaged out of equivocal data. Some details from our example will reinforce this proposal. Figure 4 is the figure provided in order to illustrate the correspondence between the Laplacian of Gaussian model and the neural data (Marr and Ullman 1981:165; Marr 1982:65). If one examines the average neural traces depicted here, and in addition the data presented in the original neurophysiology papers from which these examples were taken (Rodieck and Stone 1965: Figures 1 and 2; Dreher and Sanderson 1973), it is striking that there is a pattern of the neural response that goes un-noted by Marr and is not captured by the model – the asymmetry of peak response, depending on the polarity of the visual stimulus, and whether the bar stimulus is being swept onto the neuron's receptive field, or leaving the field. For example, the first column of Figure 4 shows that a light edge on grey background generates more neuronal response than a dark edge, whereas the model response is exactly equal. The general point is that the positing of an analogy – here that the same pattern of activation occurs in the model as for the neurons – requires selective attention to certain similarities, and the ignoring of dissimilarities. This is a matter of judgment of the scientist, and the data themselves do not usually, by themselves, force one interpretation over all others – Marr *could* have taken the asymmetry to be a relevant part of the neuronal behaviour, and come up with a mathematical model that captured this. One should not think of the structure described in any particular model as simply duplicating a structure that is pre-existing in nature.

Formal idealism does not suppose that the finding of structure in a target of investigation is purely "made up" and then projected onto the data, but takes it to be the result of the researcher's experimental interaction with the target, such that the human-dependent element of the structure can never be fully removed. One might be reminded of the way that the visual system finds shapes in what might appear as very disordered stimuli, as demonstrated with certain images in Gestalt psychology. While visual Gestalts are usually formed involuntarily, I emphasise that the scientist has a certain amount of latitude and choice in the determination of the patterns which are the target of modelling, because these depend on methods of data collection, data processing (at minimum, averaging) and style of representation.

Another way of describing the difference between formal realism and idealism, is that in the first case the abstractions of computational neuroscience are presented as if the work of the researchers has been to pare away all the extraneous neurobiological details, in order to find the essence (form) of the brain qua information processor. This is something like picking all the leaves off a tree and asserting that the bare trunk and branches are the essential structure of the tree. In contrast, the formal idealist does not assert that the computation described in the model is an essential feature of the neural circuit. The abstractions introduced by the model are taken to be there for the convenience of the scientist (i.e. to provide an economical representation which does not overload the scientist with a million details), rather than a means by which the true structures of the brain are revealed. A botanist would not insist that the leafless form is the essential structure of a tree, given the importance of the leaves in the life of the tree; nonetheless, a pared down representation would be useful, and good enough, for many purposes.

2.3 Why Formal Idealism?

Formal idealism is a doctrine of restraint: it forces one to be agnostic in response to the question of whether the brain *really is* a computer, calculating functions to which the scientist's models are a closer or wider approximation. But one must acknowledge that the literal interpretations of computational models offered by formal realism are particularly tempting in neuroscience. In other disciplines, like physics and chemistry, non-literal interpretations of computational models are more the norm. Canguilhem (1963:514-515) notes how in physics the analogical use of mathematical models does not invite one to project the ontology of the analogue-target on to the analogue-source, a caution that is often lacking when such models are used in biology. His point is that the use of an inorganic system as the analogue source for an organic target carries with it a promise of a reduction of the organic to the inorganic – i.e. the making sense of the organic in perspicacious physical terms – which is why the literal interpretations are so alluring. Canguilhem goes on to say that cybernetic models are a good example of this tendency, especially when the models' actions (e.g. in a robot), tends to simulate or mimic natural behaviour.[30] In other words, formal realism offers the promise that it is possible to devise quantitative, formal, and perspicacious models for whatever it is that the nervous system does. When this interpretation holds sway, there is a tendency to downplay the disanalogies between brains and man-made computational systems (even if the official doctrine is that the brain is *not* like a PC), and to keep the details relegated to "mere metabolic support" on the sidelines of neuroscientific investigation. This may well be limiting progress in understanding the neural basis of mental life.

The neurophysiologist Lord Adrian (1954) once quipped that, "[w]hat we can learn from the machines is how our brains must differ from them."[31] One very significant point of difference is that the hardware of electronic computers is engineered *not* to undergo material changes with use, whereas it is there is an inherent tendency for biological cells, whose material constitution is changing as they metabolise, to undergo use-based plasticity (Chirimuuta 2017; Godfrey-Smith 2016). Thus it should not surprise us that the plasticity shown by the brain, with ordinary development and deliberate learning is very much unlike what is seen in computational machines, even in artificial neural networks designed to simulate synaptic plasticity (Lake et al.

[30] "Despite their great degree of mathematical complexity, it does not appear that cybernetic models are always safe from this accident. The magical aspect of simulation is strongly resistant to the exorcism of science." Canguilhem (1963:515); Cf. Dreyfus (1972: 79-80).
[31] Quoted approvingly by Canguilhem (1963:516).

2017). The usefulness of engineering-analogues for understanding the "principles of neural design" (Sterling and Laughlin 2016) is tempered by the way that they impose an engineer's template in which structure-function relationships are fixed and transparent, and where use-dependent change is conceptualised as perturbation demanding mitigation, not a background fact of life. It could be that this very basic difference between organic and artefactual intelligence is one of the reasons why expert systems in AI, impressive as they are, have so far not made steps towards generalisation.[32]

3. CODA: LEIBNIZ THE BIOLOGIST[33]

An important supplement to the observations offered above, of Leibniz as an inventor of the computational theory of mind, is to note his views on the difference between man-made machines and living beings. He held that organic bodies were machines, but ones of infinite complexity. For unlike inorganic artefacts, the component parts of animal machines are themselves machines, and the parts of those smaller machines are also machines, *ad infinitum*.[34] Leibniz was inspired here by the recent discoveries of microscopists (Cassirer 1950), and his picture of living systems as comprising tiny machines telescoped one inside the other is not so different from that of a contemporary biologist.

I have argued in this paper that computational models, which take the workings of neural systems to be essentially like those of man-made devices -- thus rejecting Leibniz's distinction between "divine machines" and human built ones -- have been so useful to neuroscientists precisely because they remove from consideration the levels of complexity that Leibniz took to be crucial to the workings of nature. It remains to be seen whether the mysteries of biological cognition will open up to one which takes organic intelligence on its own terms. But the replacement of formal realism with an approach which pays attention to the various modes of analogy and disanalogy between brains and computers, will at least help scientists and philosophers avoid the missteps encouraged by overreaching, literal interpretations.

[32] Of course other disanalogies are most likely relevant here: e.g. embodiment of organic intelligence, whereas most expert systems are disembodied, not capable of acting in the physical world. But note that embodied AI systems (e.g. autonomous cars) have also proved to be limited in their operation outside of controlled conditions, suggesting that embodiment by itself doesn't overcome the obstacles to creating a general AI.

[33] Of course this label is anachronistic. The word "biology" was first used in 1766, fifty years after the death of Leibniz (Smith 2011: 1).

[34] As Smith (2011: 100) relates, "the animal body is not a 'mere' machine but a special kind of machine, a 'more exquisite' or 'more divine' machine, as Leibniz puts it. This is the machine of nature, or the organic body, whose exquisiteness resides in the fact that it remains a machine in its leas parts, which is to say that there is no stage in its decomposition at which one arrives at nonmachinic components."

REFERENCES

Adrian, E. D. . 1954. ' Address of the President Dr E. D. Adrian, O.M., at theAnniversary Meeting, 30 November 1953', *Proceedings of the Royal Society of London, B*, 142: 1-9.

Aizawa, Kenneth. 2018. 'Multiple realization and multiple “ways” of realization: A progress report', *Studies in History and Philosophy of Science*, 68: 3-9.

Allen, Colin, Marc Bekoff, and George Lauder (ed.)^(eds.). 1998. *Natures Purposes: Analyses of Function and Design in Biology* (MIT Press: Cambrdige MA).

Anderson, James A., and Edward Rosenfeld (ed.)^(eds.). 1998. *Talking Nets: An Oral History of Neural Networks* (MIT Press: Cambridge, MA).

Arbib, Michael A. . 2016. 'Afterword: Warren McCulloch’s Search for the Logic of the Nervous System.' in Warren S. McCulloch (ed.), *Embodiments of Mind* (MIT Press: Cambridge, MA).

Bartha, Paul. 2016. "Analogy and Analogical Reasoning." In *The Stanford Encyclopedia of Philosophy*.

Boyle, Matthew. manuscript. 'Kant’s Hylomorphism'.

Bullock, Theodore H., Michael V. L. Bennett, Daniel Johnston, Robert Josephson, Eve Marder, and R. Douglas Field. 2005. 'The Neuron Doctrine, Redux', *Science*, 310: 791-3.

Burnyeat, M. F. . 1992. 'Is an Aristotelian Philosophy of Mind Still Credible (A Draft).' in Martha C. Nussbaum and Amélie Oksenberg Rorty (eds.), *Essays on Aristotle's de Anima* (Oxford University Press: Oxford).

Canguilhem, Georges. 1963. 'The Role of Analogies and Models in Biological Discovery.' in A. C. Crombie (ed.), *Scientific Change* (Basic Books: New York).

———. 2008. 'Machine and Organism.' in Paola Marrati and Todd Meyers (eds.), *Knowledge of Life* (Fordham University Press: New York).

Cao, Rosa. 2014. 'Signaling in the Brain: In Search of Functional Units', *Philosophy of Science*, 81: 891-901.

Cassirer, E. 1950. *The Problem of Knowledge: Philosophy, Science, and History since Hegel* (Yale University Press: New Haven).

Chirimuuta, M. 2017. 'Crash Testing an Engineering Framework in Neuroscience: Does the Idea of Robustness Break Down?', *Philosophy of Science*, 84: 1140–51.

———. 2018a. 'Explanation in Computational Neuroscience: Causal and Non-causal', *British Journal for the Philosophy of Science*, 69: 849 - 80.

———. 2018b. 'Marr, Mayr, and MR: What functionalism should now be about', *Philosophical Psychology*, 31: 403-18.

———. forthcoming. 'Charting the Heraclitean Brain: Perspectivism and Simplification in Models of the Motor Cortex.' in Michela Massimi and Casey McCoy (eds.), *Understanding Perspectivism: Scientific Challenges and Methodological Prospects* (Routledge: New York).

Craver, C.F., and Lindley Darden. 2013. *In Search of Mechanisms* (Chicago University Press: Chicago, IL).

Craver, C.F., and David Michael Kaplan. 2018. 'Are More Details Better? On the Norms of Completeness for Mechanistic Explanations', *British Journal for the Philosophy of Science*.

Craver, C.F., and James Tabery. 2017. "Mechanisms in Science." In *The Stanford Encyclopedia of Philosophy*.

Dardashti, R., Karim P. Y. Thébault, and Eric Winsberg. 2017. 'Confirmation via Analogue Simulation: What Dumb Holes Could Tell Us about Gravity', *British Journal for the Philosophy of Science*, 68: 55-89.

Daugman, John G. 2001. 'Brain Metaphor and Brain Theory.' in William Bechtel, Pete Mandik, Jennifer Mundale and Robert S. Stufflebeam (eds.), *Philosophy and the Neurosciences: A Reader* (Blackwell: Oxford).

Davis, Martin. 2000. *The Universal Computer: The Road from Leibniz to Turing* (W. W. Norton & Company: New York).

Dreher, B., and K. J. Sanderson. 1973. 'Receptive Field Analysis: Responses to Moving Visual Contours by Single Lateral Geniculate Neurones in the Cat ', *J. Physiology*, 234: 95-118.

Dreyfus, Hubert L. 1972. *What Computers Can't Do: A Critique of Artificial Reason* (Harper & Row: New York).

Egan, Frances. 2017. 'Function-Theoretic Explanation and the Search for Neural Mechanisms.' in David Michael Kaplan (ed.), *Explanation and Integration in Mind and Brain Science* (Oxford University Press: Oxford).

Elgin, Catherine Z. 2004. 'True Enough', *Philosophical Issues*, 14: 113-31.

Fairhall, Adrienne. 2014. 'The receptive field is dead. Long live the receptive field?', *Current Opinion in Neurobiology*, 25: ix–xii.

Frégnac, Yves. 2017. 'Big data and the industrialization of neuroscience: A safe roadmap for understanding the brain?', *Science*, 358: 470-77.

Godfrey-Smith, P. 2016. 'Mind, matter, and metabolism', *Journal of Philosophy*, 113: 481–506.

Goldstein, K. 1938. *The Organism: A Holistic Approach to Biology Derived from Pathological Data in Man* (American Book Company: New York).

Jonas, E., and K. Kording. 2017. 'Could a neuroscientist understand a microprocessor?', *PLoS Comput. Biol.*: e1005268.

Kant, Immanuel. 1929. *The Critique of Pure Reason* (Palgrave: Basingstoke).

Kaplan, David Michael. 2011. 'Explanation and Description in Computational Neuroscience', *Synthese*, 183: 339-73.

Kline, Ronald R. 2015. *The cybernetics moment: or why we call our age the information age* (John Hopkins University Press: Baltimore, MA).

Knuuttila, Tarja, and Andrea Loettgers. 2014. 'Varieties of noise: Analogical reasoning in synthetic biology', *Studies in History and Philosophy of Science*, 48: 76-88.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. 'Building machines that learn and think like people', *Behavioral and Brain Sciences*, 40.

Lettvin, Jerome. 2016. 'Foreword to the 1988 Reissue.' in Warren S. McCulloch (ed.), *Embodiments of Mind* (MIT Press: Cambridge, MA).

Lettvin, Jerome Y., H. R. Maturana, Warren S. McCulloch, and Walter H Pitts. 1959. 'What the Frog's Eye Tells the Frog's Brain', *Proceedings of the IRE*, 47: 1940-59.

Longuenesse, Béatrice. 2005. *Kant on the Human Standpoint* (Cambridge University Press: Cambridge).

Mante, Valerio, David Sussillo, Krishna V. Shenoy, and William T. Newsome. 2013. 'Context-dependent computation by recurrent dynamics in prefrontal cortex', *Nature*, 503: 78-84.

Marcus, Gary. 2015. 'The Computational Brain.' in Gary Marcus and Jeremy Freeman (eds.), *The Future of the Brain* (Princeton University Press: Princeton, NJ).

Marr, David. 1982. *Vision* (W. H. Freeman: San Francisco).

Marr, David, and Shimon Ullman. 1981. 'Directional selectivity and its use in early visual processing', *Proceedings of the Royal Society of London, B*, 211: 151-80.

Mayr, Ernst. 1988. 'The Multiple Meanings of Teleological.' in Ernst Mayr (ed.), *Toward a New Philosophy of Biology* (Belknap Press of Harvard University Press: Cambridge, MA).

McCulloch, Warren S., and Walter Pitts. 1943. 'A Logical Calculus of the Ideas Immanent in Nervous Activity', *Bulletin of Mathematical Biophysics*, 5: 115-33.

Miłkowski, Marcin. 2018. 'From Computer Metaphor to Computational Modeling: The Evolution of Computationalism', *Minds and Machines*.

Morar, Florin-Stefan. 2015. 'Reinventing machines: the transmission history of the Leibniz calculator', *British Society for the History of Science*, 48: 123-46.

Nussbaum, Martha C., and H. Putnam. 1992. 'Changing Aristotle's Mind.' in Martha C. Nussbaum and Amélie Oksenberg Rorty (eds.), *Essays on Aristotle's de Anima* (Oxford University Press: Oxford).

Papert, Seymour. 2016. 'Introduction.' in Warren S. McCulloch (ed.), *Embodiments of Mind* (MIT Press: Cambridge, MA).

Piccinini, Gualtiero. 2004. 'The First Computational Theory of Mind and Brain: A Close Look at Mcculloch and Pitts's "Logical Calculus of Ideas Immanent in Nervous Activity"', *Synthese*, 141: 175-215.

Pickering, Andrew. 2010. *The Cybernetic Brain: Sketches of Another Future* (Chicago University Press: Chicago, IL).

Polger, Thomas W., and Lawrence A. Shapiro. 2016. *The Multiple Realization Book* (Oxford University Press: Oxford).

Potochnik, Angela. 2017. *Idealization and the Aims of Science* (Chicago University Press: Chicago, IL).

Putnam, H. 1988. *Representation and Reality* (MIT Press: Cambridge, MA).

Ritchie, J. Brendan, and Gualtiero Piccinini. 2018. 'Computational Implementation.' in Mark Sprevak and Matteo Colombo (eds.), *The Routledge Handbook of the Computational Mind* (Routledge: London).

Rodieck, R. W., and J. Stone. 1965. 'Response of Cat Retinal Ganglion Cells to Moving Visual Patterns', *Journal of Neurophysiology*, 28: 819 - 32.

Rosenblueth, Arturo, Norbert Wiener, and Julian Bigelow. 1943. 'Behavior, Purpose and Teleology', *Philosophy of Science*, 10: 18-24.

Seidengart, Jean. 2012. 'Cassirer, Reader, Publisher, and Interpreter of Leibniz's Philosophy.' in R. Kroemer and Y. C. Drian (eds.), *New Essays in Leibniz Reception: In Science and Philosophy of Science 1800-2000* (Springer: Basel).

Shagrir, Oron. 2010. 'Brains as analog-model computers', *Studies in History and Philosophy of Science*, 41: 271-79.

———. 2018. 'The Brain as an Input–Output Model of the World', *Minds and Machines*, 28: 53-75.

Shenoy, K. V., M. Sahani, and M. M. Churchland. 2013. 'Cortical Control of Arm Movements: A Dynamical Systems Perspective', *Annual Review of Neuroscience*, 36.

Smith, Justin E. H. . 2011. *Divine Machines: Leibniz and the Sciences of Life* (Princeton University Press: Princeton, NJ).

Sterling, P., and S. Laughlin. 2016. *Principles of Neural Design* (MIT Press: Cambridge, MA).

Walshe, Francis M. R. 1951. 'The Hypothesis of Cybernetics', *British Journal for the Philosophy of Science*, 2: 161-3.

———. 1961. 'Contributions of John Hughlings Jackson to Neurology', *Archives of Neurology*, 5: 119-31.