Mechanistic Philosophy and the Use of Deep Neural Networks in Neuroscience

Exam Number: B110919

Word Count: 8,500[1]

MSc Taught Philosophy 2017/2018

[1] With permission from the post graduate office

Acknowledgments

I am indebted to Julian Hausser, Nina Poth, George Deane, Matt Sims, Luke Kersten, Liv Coombes, Jan Meyer, and Katherine Harrison whose feedback, discussion, and support has been an invaluable aid in furthering the development of this paper.

I also thank the staff at Peter's Yard, whose constant and unbroken supply of free coffees have partially gone into making this paper what it is.

Finally, none of this would have been possible without the help and support of my supervisors, Dr. Mark Sprevak and Dr. Joseph Dewhurst.

Abstract

Deep Learning has revolutionized artificial intelligence (AI) over the past decade (LeCun et. al. 2015). This has led to the creation of 'neurally inspired' *deep-neural-networks* (DNNs). DNNs are claimed to be biologically realistic in the sense of incorporating key mechanistic or architectural features of the brain, such as a hierarchical structure. Interestingly, they are also said to exhibit similar behavioral capacities, as they are capable of performing 'near human-level' on a variety of behavioral tasks (Kriegeskorte 2015). These similarities have led researchers to propose DNNs as biologically realistic models of behavior and brain function (ibid). In this paper, I argue that there are at least two concerns in relating DNNs to the brain. First, I suggest that DNNs might not, as they stand, exhibit biologically realistic behavior. Secondly, I argue that there are key mechanistic dissimilarities between biological and artificial neural networks that impedes a so-called *model-mechanism-mapping* relationship. Thus there is a mismatch between the two systems at both the *behavioral* and *implementational* levels. The explanatory status of DNNs is accordingly called into question. To defend this position, I make recourse to the *mechanistic philosophy of science*, a framework within which to assess a model's explanatory status (Craver 2007).

Keywords: mechanistic philosophy | deep neural networks | models | computational neuroscience| behavior

1. Introduction

A central aim of neuroscience is to explain how flexible behavior and intelligence arise from a tangle of neural circuits and biochemical activities (Fregnac 2017). But the brain is a highly degenerate, nonlinear, and hierarchical system, and thus relating behavior and mentation to neural activities remains one of the most difficult questions in neuroscience today. One way to provide an explanation is to focus on higher-level conceptual work to frame lower-level research into mechanistic underpinnings and their global organization (Lake et. al. 2017; Churchland, Sejnowski 2016; Sejnowski et. al. 2014). To paraphrase the mechanistic philosopher Bechtel (2009), explaining a biological system requires not only looking 'down' and 'around' at organized parts and activities, but also 'up' at the environmental context in which those mechanisms are embedded. To capture multi-level explanations requires more than conceptual work, however. Researchers additionally require computational models

to assist in their comprehension of a system (Bechtel, Levy 2013). Interestingly, due to a recent technological shift—a sea-change known as the rise of big-data—computational models have become an integral tool in almost all areas of neuroscience. Of particular note are *deep-neural-networks* (DNNs; Hinton 2007; LeCun, Bengio, Hinton 2015).

Like their computational predecessors, DNNs are situated in McClelland & Rumelhart's *parallel distributed processing* framework (1987). But in contrast to earlier *artificial-neural-networks* (ANNs), DNNs signify a marked improvement in computational power, surpassing all previous structures on historically challenging problems, such as object recognition (Yamins et. al. 2014; Yamins, DiCarlo 2016b; Lehky, Tanaka 2016), machine vision (Majaj, Pelli 2018), machine translation (Sutskever et. al. 2014), and even speech recognition (Sak et. al. 2014).

How and why are DNNs different? First, ANNs traditionally consist of *input-output layers* that interface with the world and emit a behavioral response, respectively. Sandwiched between the two are a series of computational *units* distributed throughout a *hidden layer*. This layer is responsible for encoding, transforming, and propagating information to the output layer. DNNs build on this structure, but in addition use a powerful technique called *deep-learning* that involves distributing units throughout multiple (hidden) layers, leading to the characteristically deep and hierarchical architecture that dominates the field today (Kriegeskorte 2015). Additionally, researchers in the field commonly invoke the biological inspiration behind DNNs, citing not only architectural similarities, but similar computational features such as the development of *receptive-fields* (Yamins, DiCarlo 2016a; but see Dong, Wang, Hu (2018)). One exciting development that emerges from this is the potential use of DNNs as models for animal behavior and neural information processing (Kietzmann et. al. 2017; Glaser et. al. 2018; Kriegeskorte 2015; Marblestone et. al. 2016; Hassabis et. al. 2017; van Gerven 2017; Helmstaedter 2015). Are DNNs, then, a promising inroad into explaining biological systems?

The thesis defended in this paper is that DNNs *can* facilitate in the process of explaining behavior and brain function, but that, as they stand, there are a couple conceptual difficulties in calling them explanatory of biological systems (Carlson et. al. 2017; Kay 2017). Moreover, I approach this topic using *mechanistic philosophy of science*, which emphasizes the importance of properly *characterizing* the *explanandum*—the behavior or phenomena to be explained—as well as relating a model to its target system by identifying a minimum set of similarity relationships (Craver 2007). Essentially, this recapitulates debates in the modeling literature concerning *biological realism*, that is, the biological fidelity of our models (Crick 1989). Put forth as a question, what does the mapping relationship have to be between DNNs and brains for the former to explain the latter? What equivalencies can we identify

between both systems? Do the two systems, in fact, *behave* the same way or *solve* the same tasks? Ultimately, I will argue that, as they stand, DNNs might not map onto their target system at (at least) two levels: the behavioral (Lake et. al. 2017), and the implementational (Brette 2015). That is, despite their 'near human-level' performance, the two systems might not only operate (mechanistically) in different ways but, might *behave* in relevantly dissimilar ways. In each case, the putative mapping relationship becomes slightly skewed and the explanatory status of current DNNs is called into question.

The goal of this paper is not to criticize DNNs, or deflate their instrumental value in neuroscience. Rather, it is to motivate directing our attention to the precise specification of the *behavior* of biological systems, a step that is often treated as a mere preliminary endeavor (Shagrir, Bechtel 2017). Only after properly characterizing behavior can we determine the explanatory status of a model thereof. Without this first step, we might be tacitly guided by (sometimes outmoded) assumptions dwelling at the behavioral-level (Edelman 2016). I also aim to suggest that, according to the mechanistic framework, it is ultimately the implementational features that *explain*. Therefore, merely reaching similar performance benchmarks does not make a model explanatory.

The structure of this essay will be as follows. In section (2), I introduce the mechanistic philosophy of science and assess whether or not DNNs—as computational models—are amenable to this framework. Section (3) addresses the equivalencies of DNNs and the brain/organisms by looking at the behavior of the two systems. Section (4) proceeds to the mechanistic level in order to address the implementational (dis)similarities between the two systems. In particular, I look at the neural codes employed in both biological and artificial neural networks. Section (5) concludes by suggesting that, while perhaps not satisfying similarity relationships at both levels, DNNs still function as invaluable tools in neuroscientific research.

2. The Mechanistic Framework

The mechanistic framework is explicitly concerned with questions of biological realism in relating a model to its target system (Kaplan, Craver 2011). This section attempts to sketch a couple key features of both *mechanistic explanations* (MEs) and *mechanistic models*, before addressing whether DNNs are properly construed as mechanistic. With a conceptual framework on the table, we can proceed to see how DNNs align (or not) with desiderata presented herein on both the *behavioral* (section 3) and *implementational* (level 4) levels.

The main aim of mechanistic philosophy of science is to assess how a phenomenon of interest, φ, is situated in the causal structure of the world (Salmon 1984). In other words, to explain φ requires elucidating the causally relevant features that realize it. Often through *decomposition* and *localization* techniques (Craver 2007)[2]. Examples of φ include the mechanism for protein synthesis, the generation of the action potential, and neurotransmitter release. Consider the latter. Neurotransmitter release (NTR) is the process by which an action potential depolarizes the axon terminal and eventuates in the release of transmitter across the synaptic cleft (ibid.). As Machamer et. al. note, φs are sometimes demarcated by 'start' and 'finish' conditions (2000). In this case, depolarization initiates the process, and NTR terminates upon the release of transmitter. Between both start and finish there are biochemical cascades mediating the process, such as calcium ionic influx (which sustains NTR) and the docking of synaptic vesicles to the membrane.

While truncated for simplicity, the above example adumbrates two key desiderata governing MEs. First, and most importantly, is the specification of the *explanandum* φ (e.g., NTR). The *explanandum* assumes a central position as it ultimately frames notions of causal relevance. Stated differently, it is only in light of the φ that one may reliably infer whether an entity or activity is causally implicated in producing it. This is because mechanisms are always mechanisms *of* a phenomenon (Glennan 1996). MEs require more than the specification of parts and activities, however. One must also show how they are organized so as to produce φ.

The second desideratum is that to explain φ requires the identification and inclusion of *only* causally relevant details. To explain φ, then, requires showing how causally relevant entities and activities underlie it. For example, it is known that calcium is causally relevant for NTR; conversely, sodium, which enters the cell contemporaneously, is causally *irrelevant* (Craver 2007). The question, of course, becomes: how does one determine the causal status of a given entity or activity? While this is a contentious question, one way to address it is the *causal interventionist approach* (Woodward 2003, 2017). The interventionist holds that researchers can reliably infer the relevance of an entity if they can manipulate or intervene on its activities and observe downstream effects (ibid.). In the above example, researchers Katz & Miledi (1967) determined the causal relevance of calcium influx by manipulating extracellular concentration and noting that upon eliminating it from the solution NTR no longer

[2] Briefly, *decomposition* is the process by which researchers break φ down into smaller functions and subfunctions; *localization* is the process by which researchers localize and identify mechanisms that produce φ.

occurred. By similar means, they determined that sodium is irrelevant in the process, as eliminating it from the solution did not compromise the eventuation of NTR. While the interventionist approach faces the challenge of differentiating causal from correlational effects (Jazayeri, Afraz 2017), it is one principled inroad into the problem.

An important distinction that falls out of the above discussion is the difference between *how-possibly* and *how-actually* explanations. The former are said to have 'explanatory purport' but are only loosely constrained by facts about the system. Thus, they are not 'fully' explanatory (Craver 2007). Moreover, they typically gloss over over some or most of the mechanisms implicated in producing φ. The latter, conversely, qualify as complete explanations. That is, they "represent all and only the *relevant* portions of the [explanandum]" (ibid.: 27). Craver (ibid.) suggests that the lacunae in our MEs are slowly filled until we 'complete' the process, completeness always being relative to φ (Milkowski 2016).

It is often the case, however, that researchers do not have direct access to the mechanisms underpinning φ, and must resort to *model building* to capture its behavior and organization (Bechtel 2008; Zednik 2015; Milkowski 2013; Matthewson 2017a, 2017b). While there are a variety of model types (Bassett, Zurn, Gold 2018), *mechanistic models* are our primary concern. There are two corresponding desiderata in determining a mechanistic model's explanatory status. For one, it is critical to delineate the explanandum of the system under investigation, "otherwise the model's use and value will be unclear" (Milkowski 2016: 1460). Following Glennan (2005), I call this the *phenomenal description* of the system. Secondly, mechanistic models evince how the system is situated in the causal structure of the world. It does this by representing the target system's causally relevant activities and entities. This is called the *mechanistic description* (ibid.).

The latter desideratum is often called a *model-mechanism-mapping* (3M) relationship (Kaplan 2011). That is, a mechanistic model is explanatory if (1) variables in the model correspond to entities and activities in the target system; and (2) the dependencies posited among the variables equate to causal relations between the components and activities of the mechanism (ibid.). Similar to MEs, mechanistic models are subject to completeness criteria. For instance, Craver suggests that there are *how-possibly* or *how-plausibly* models, which suggest how a system or mechanism works, but are only loosely constrained by biophysical considerations (Craver 2007). Similarly, these models often 'black-box' the mechanisms that produce the phenomenon. Craver argues that a model becomes more explanatory as these black-boxes are 'filled' and the model transitions towards a *how-actually* model,

one that represents only *relevant* portions of the target system (ibid.). It will be important to keep this distinction in mind as we proceed to our discussion of DNNs.

A model thus has to satisfy at least two similarity relationships to qualify as a mechanistic model of its target system: the *phenomenal description*, which characrterizes its behavior; and the *mechanistic description*, which focuses on its causal features. While the former is critical for guiding research, it is ultimately the latter that explains: "while a phenomenal description… is useful for describing and predicting the mechanism's behavior, it is the mechanistic model… that actually explains" (Glennan 2017a: 66). The above is a general account of MEs and mechanistic models, and more could be said (cf. Craver 2007). Nevertheless, I take the preceding as establishing the core of the mechanistic framework.

It might be suggested, however, that the mechanistic framework has no obvious connection to computational models, especially those situated in the network approach such as ANN/DNNs (cf. Chirimuuta 2014). Indeed, it could be argued that DNNs—comprised of homogenous and simplistic *units*—cannot be *decomposed* or *localized* as the standard mechanistic arguments suggest (Silberstein, Chemero 2013). For instance, units are functionally indistinguishable and hence interchangeable, and thus do not lend themselves to procedural methods such as decomposition, at least not in the same respect as biological neural networks (ibid). This is a common argument and might evince the limitations of paradigmatic mechanistic approaches. Yet it is unclear whether decomposition and localization exhaust the tools of the mechanistic framework (2014, 2018). For instance, equally integral to the mechanistic framework is the *interventionist approach*, and in this case researchers often manipulate or intervene in their networks in ways similar to studies performed on organisms.

Interestingly, Kietzmann et. al. argue that the interventionist approach allows researchers to assess the relevance of architectural features (feedforward or recurrent structures, say), connection weights, additional layers, network statistics, and so on (2017: 14). As Zednik puts it:

> Networks models… emphasize organization properties, individual component parts, and operations, while still rendering the relevant network mechanism amenable to interventions… and control. Insofar as this is one of the principle norms of [MEs], models that describe brain networks in ways that satisfy this norm should be considered mechanistic. (2014: 17)

Moreover, while the above criticisms might apply to traditional ANN approaches, there has been a recent development in machine-learning called *explainable AI* (Bornstein 2016). Largely due to rising

concerns of inscrutability and incomprehensibility—the so-called 'Bonini's paradox'[3]—researchers have started developing techniques for *visualizing* the activation patterns of individual units, so as to render their activities more interpretable (Yosinski et. al. 2015; Doshi-Veliz, Kim 2017). Thus, the means of decompositions and localization, commonly seen as proprietary to the biological sciences, seem to be encroaching upon the DNN community, since they offer a new heuristic for surmounting inscrutability claims (Bornstein 2016).

For the rest of this paper, I take DNNs as amenable to mechanistic analysis (for more detail, cf. Piccinini, Craver 2012; for contrast, cf. Ross 2015). With the preceding on the table, we may better approach the question at hand: do DNNs satisfy key mechanistic desiderata specified herein? That is, do they exhibit the phenomenal description or the task to be explained (e.g., object or speech recognition)? Do they reference causally relevant details at some relevant level of finer grain? In the rest of this paper, I set out to suggest two problems facing a mapping relationship between DNNs and the brain.

3. The Phenomenal Description: A Comparison at the Behavioral-Level

In debates concerning biological realism, the questions almost invariably address the mechanistic components of a model. However, it is also important to look up to assure that the behavior of the model is itself biologically realistic. In mechanistic terms, this amounts to specifying the *explanandum* or phenomenon to be explained. It is thus important to have a good sense of the system's behavior, as it is only against the explanandum that a model can be assessed. This section, then, addresses the relationship between DNNs and brains/animals on the *behavioral-level* to see if the two exhibit comparable *phenomenal descriptions*.

To begin, what does it mean, to behave? The answer seems self-evident: behave is what animals *do* (Edelman 2016). As it turns out, characterizing behavior is one of the most important, if often underestimated, steps in explaining biological systems, as choosing among the many things animals *do* influences the models and explanations we produce (Kaufmann 1971). Nonetheless, perhaps due to its *prima facie* self-evidence, the behavioral-level is particularly prone to assumptions and biases that we typically pass over in favor of mechanistic underpinnings. As Shagrir & Bechtel note: "the specification of phenomena is generally treated as unproblematic—the problem is

[3]Bonini's Paradox concerns the construction of a model that is as complex as the target system being modeled, thereby compromising its explanatory potential.

explaining them" (2017: 190). But rather than a preliminary endeavor, characterizing behavior should be seen as something fundamental in its own right (Krakauer et. al. 2017). Thus, in this section I assess whether DNNs (1) engage in the same *learning tasks* and (2) *behave* in the same way. In particular, it will be argued that by specifying behavior in terms of the information processing tasks being performed (typically in terms of input-output profiles), and not factoring in other features (an animal's learning tasks or intrinsic motivation (Baldasarre 2011)), we are led to perceive a commonality between the two systems (Edelman 2016). It will be argued that by updating our conception of behavior, a mapping relationship between DNNs and animals becomes more problematic.

3.1 Is 'Near Human-Level' Near Enough?

In the introduction I addressed a couple factors behind the DNN success story. While work in deep-learning required notable theoretical work, it was largely advancements in computational power that accelerated DNN research. As Hinton jocularly remarked, it took 17 years to get deep-learning right: one year thinking, and 16 years of progress in computing (quoted in Cox, Dean 2014). One such development was the unprecedented access to petabytes of (un)labeled data from the internet ('big-data'). Perhaps, however, this is where we can see our first palpable difference between DNNs and animals: the former require ponderously large datasets and (mostly) supervised learning (Glaris et. al. 2018); the latter are capable of learning from few examples and operate according to internally driven cost functions that are not set or supervised explicitly by the environment, as feedback is often sparse through time (Marblestone, Wayne, Kording 2016).

More perspicuously, the two—DNNs and animals—seem to engage in different *learning tasks*, suggestive of a wide dissimilarity in terms of the goals they achieve. Indeed, Lake et. al. motivate similar concerns in their influential paper, "Building machines that learn and think like people" (2017). They argue that despite impressive near human-level performance on a variety of behaviorally relevant challenges (e.g., character recognition (Ciseran et. al. 2012)), the two systems might in fact achieve their goals by solving different learning task. To illustrate this point, consider the recent excitement surrounding AlphaGo's win against world champion Lee Sedol at the game of Go (Silver et. al. 2016). Before competing in its first game, AlphaGo was trained on 28.4 million possible board positions. It subsequently played an initial batch of 30 million games before continuing to play successively stronger versions of itself. By the time of its match with Sedol, AlphaGo completed 100 million games in total. In striking contrast, Sedol is estimated to have played around 50,000 games in his entire life. Considering such a wide disparity, it is surprising that Sedol could keep up at all (Lake et. al. 2017).

This example is both impressive and illustrative. Impressive, as it evinces the computational power of current deep-learning techniques as embodied in DNNs; illustrative as it indicates that AlphaGo, despite reaching and then surpassing expert human-level performance, might perform different learning tasks to reach its goal: "it is not that [DNNs] and people are solving the same tasks differently. They may be better seen as solving different tasks" (ibid.: 9). The justification for this remark is that humans, and animals more generally, approach novel situations with an extensive backdrop of prior experience. As a result, humans have an edifice of domain-general and domain-specific knowledge before they even begin.

Additionally, human cognition seems to consist of—and be bootstrapped by—several cognitive 'ingredients' (i.e., knowledge-transfer and learning-to-learn) and developmental 'software' (i.e., intuitive psychology; ibid.: 9-19). Conversely, DNNs embody a strong empiricist conception of the world reminiscent of Turing's blank notebook[4]. Stated differently, they begin with little to no prior knowledge, and thus learn everything from scratch (hence their need for laborious training and large datasets). From this, Lake et. al. draw the following conclusion: "Human learners fundamentally take on different learning tasks than today's neural networks, and if we want to build machines that learn and think like people, our machines need to confront the kinds of tasks that human learners do" (ibid.). A guiding question, then, is: what would it take for AlphaGo to achieve human-level performance with as (comparatively) little data as Sedol? Lake et. al. propose that this can be done by approximating human learning tasks more closely. Not only would this potentially augment DNN performance, but perhaps unravel the mysteries of human cognition (ibid.: 23).

Lake et. al.'s prescient analysis conveys the need for a strong conceptual framework. By carefully specifying the learning tasks humans and other animals perform, it becomes clear that DNN's impressive near human-level performance is not enough to establish a similarity between model and target system at the behavioral-level. Until DNNs approximate at least some aspects of the *phenomenal description*, however, the amount of explanatory insight they provide is questionable. This is not to say they cannot achieve this, just that in assessing the explanatory status of a model it is paramount that the two systems exhibit the same *phenomenal description*, which, in turn, requires carefully specifying what the system does.

[4] Turing depicted a child's mind as similar to a blank notebook with "rather little mechanisms and lots of blank sheets", with the pages slowly filling up following experience, particularly in terms of responses to rewards and punishments, recapitulating the strong behaviorist tendencies of his time (Lake et. al. 2017).

This interlocks well with the mechanistic point of judiciously specifying the explanandum. Importantly, this goes beyond the specification of a problem in terms of input-output relations (Shagrir, Bechtel 2017). In other words, if we characterize human player's behavior or the problems they face in terms of the information they begin with (their input) and the state in which they need to end up (their output), then AlphaGo appears to achieve what humans achieve. However, while specifying an input-output profile is perhaps *one* important ingredient in characterizing the tasks an animal faces, it does not follow that a model which exhibits this input-output profile in fact *behaves* the same way. In fact, Shagrir & Bechtel argue that one trivializes the behavioral-level by looking at an animal's task merely in terms of the inputs with which it deals (2017: 207). In our above example, then, it is not just that the task consists of board positions (input) and winning (output). Rather, it is whether or not AlphaGo performs similar *learning tasks* to human agents. As suggested above, this does not appear to be the case. The upshot is that identifying comparable input-output profiles might actually mislead us into perceiving behavioral similarities between the two systems (ibid.). That is, if we conceive of AlphaGo's tasks as simply winning, then AlphaGo satisfies the *phenomenal description* of the target system. But if the task is *reconceived* as learning from a limited number of trials, say, then DNNs no longer seem to correspond.

Interestingly, the conception of behavior as a mapping from inputs to outputs has a long and complicated history in the mind and brain sciences (cf. Skinner 1957). Edelman (2016), for instance, argues that this so-called *stimulus-response arc* (often called an input-output doctrine (Thompson 2007)) operates as an implicit assumption in these disciplines. In brief, the stimulus-response arc assumes that organismic behavior consists of a set of *reflexes* or *responses* to incoming pleasant or noxious stimuli, which are either hardwired or conditioned using reinforcement learning associated with reward and punishment (Graham 2017). It has recently been argued, however, that this preconception is largely outmoded and unfounded, yet still influences the explanations and models we provide (Krakauer et. al. 2017). Given that behavioral-level assumptions frame the entire research program, it is thus "particularly important to be open to the possibility that we have them wrong" (Edelman 2016: 751). It is crucial, then, from a mechanistic perspective to subject our assumptions to critical assessment, especially since they in part determine the explanations and models we produce (ibid.). If I am right in suggesting that conceiving AlphaGo's task in terms of input-output relations leads us into perceiving a correspondence with humans, then this suggests that the stimulus-response arc still

influences research today (Edelman 2016). To elaborate, I will look at the behavior of *deep feedforward convolutional neural networks*[5] (CNNs).

CNNs have been used to model longstanding computational difficulties associated with human and animal cognition, such as object classification, shape detection, and even certain linguistic tasks (Kubilius et. al. 2016; Khaligh-Razavi, Kriegeskorte 2014; Sutskever et. al. 2014). Currently, CNNs are seen as one of the most promising means to model these behaviorally relevant tasks, since not only have they outperformed all pre-existing models to date, but also rival their target systems on object classification under normal viewing conditions (Yamins, DiCarlo 2016a). Without degrading their successes (especially given their simpler architectures), it is important to note that CNNs can be characterized in terms of the above stimulus-response arc (Edelman 2016). They exhibit this behavior in that they excel at *categorization problems*, tasks that involve mapping entire structures to entire structures in a way that resembles a Google or Watson query (ibid.). Insofar as CNNs are used to model behaviorally relevant tasks, it is important to assess whether or not they correspond to organismic behavior. This does not seem to be the case, however, as organismic behavior is *agentic* and *self-initiated* in a way that goes beyond stimulus-response contingencies, as we will see below (Krakauer et. al. 2017; Anderson & Perona 2014; Cooper & Peebles 2015; Fetsch 2016; Frank & Badre 2015; Fregnac 2017).

Despite the fields of biology and ethology proposing more ecologically nuanced conceptions of behavior (Gomez-Marin et. al. 2014), the stimulus-response arc persists not only in machine-learning, but in the mind and brain sciences (Edelman 2016). It seems to persist because it subjects the organism to an "explicitly causal analysis", one in which behavior can be captured by measurable responses to controlled stimuli parameters (Bowers 1973). Yet despite its practical utility, much is lost in casting the organism in terms of mapping inputs to outputs. For instance, this seems to obscure the *agentic* or *active* nature of behavior. Put differently, organismic activity is self-initiated "in a manner that blurs the distinction between a stimulus, which could be an act… of the agent, and a response, which could be the resulting snapshot of the environment" (Edelman 2016: 754). Furthermore, it does not

[5] The following argument should not be taken as an argument against DNNs *simpliciter*, but CNNs. DNNs come in a variety of forms and consist of subtly or dramatically different network architectures. One such network that goes beyond input-output mappings is a *recurrent neural network* (RNN), which consists of lateral and feedback connections. Recurrent topology allows RNNs to approximate dynamic systems, meaning their output is also dependent on occurrent network activity, as well as previous computations. They are not simply classifiers in the same way as CNNs. My argument should not be taken as addressing RNNs, but, for simplicity, as addressing CNNs.

do justice to the *spontaneous* activity that characterizes nervous tissue and flexible behavior, which makes it difficult to identify singular reflexes that respond exactly the same way upon even the same stimulus parameters (ibid.).

A further difficulty with the stimulus-response arc is that it indiscriminately entails that any coordinated response to observable stimuli qualifies as behavior (Krakauer et. al. 2017). This appears to be a mistake, as certain stimuli hold more relevance to the organisms than others: "animals are constantly in motion… which invites the following easy mistake: since an animal is responding to stimuli, and physiological processes are measurable, one is therefore studying an animal's behavior… [But] it is a significant confusion to label a coordinated response to a stimulus a 'behavior' without first determining the relevance of the response to the animal's life" (ibid.: 482). The problem, to state it differently, in this 'easy mistake' is that it couches behavior solely in terms of the *observable inputs* with which the animal interacts and the *measurable responses* thereby produced, but does not consider which stimuli are *relevant* to the animal's life. That is, the notion of relevance is not evident merely in the input-output relations, as it is in part determined by the *intrinsically motivated* and *agentic* features of the animal (Baldassarre 2011; Edelman 2016). Thus, as Thompson puts it: "it is crucial to distinguish between information about stimuli as defined by an observer and information in the sense of what meaning the stimuli have for the animal" (2007: 43).

The important point for present purposes is that the input-output doctrine has not only operated as a tacit preconception in machine-learning, but in the mind and brain sciences as well (Edelman 2016). I argue that this correspondingly leads us into perceiving a similarity between the two systems in which CNN models satisfy the *phenomenal description* requirement. However, if we *reconceive* behavior as something beyond stimulus-response contingencies, then the mappability of model to target system at the behavioral-level becomes more problematic. I have argued briefly above as to why organismic behavior may not be adequately conceptualized in terms of stimulus-response arcs. I now bring this back to our current evaluation of CNN models.

Consider the difficulty in treating linguistic production in terms of input-output mappings (Chomsky 1959; Edelman 2016). This has traditionally been a longstanding problem for a stimulus-response conception of behavior, largely due to its diachronicity and high-dimensional search space in which "the next word in an utterance may depend in principle on the speaker's entire past life experience, as well as another word choice that is still in the future" (ibid.: 756). Nonetheless CNNs have been used in modeling a myriad of linguistic tasks. How are CNNs capable of modeling the very behavior that seemed to challenge such an approach? As Edelman notes, this somewhat surprising

ability can be explained "if one considers closely the kinds of language tasks on which [these DNNs] do well. Invariably, these involve the mapping of entire structure to entire structures (as in learning to choose the best parse tree for a given input sentence)" (ibid.: 754). But language production, so I argue, is not adequately conceived of as a classification or mapping between a (finite) range of a large set of sentences and a large set of sentences (Marcus 2018). Rather, it is better viewed as "a potentially infinite range of input sentences and an equally vast array of meanings, many never previously encountered" (ibid.). The take away point is that by updating our conception of behavior to account for the *agentic* and *intrinsically motivated* nature of environmentally situated animals, then the conception of behavior as a stimulus-response arc begins to lose tractability. Thus, the initial grounds for perceiving a commonality between CNNs and animal behavior begins to give way: "[CNNs] can serve as models of behavior only if we pretend that behavior amounts to one Google [query] after another" (Edelman 2016: 754).

It is worth mentioning, however, that many researchers in machine-learning are now skeptical of CNNs as models of behavioral tasks, and they are by no means representative of the entire DNN wheelhouse. Nevertheless, the example of CNNs (still a prominent model choice (Kietzmann et. al. 2017)) illustrates the mechanistic point of carefully characterizing the behavior of biological systems, as this ultimately frames the models and explanations we subsequently provide. Thus, in the same way that Lake et. al. (2017) stress the importance of properly identifying the learning tasks of human agents, and therefore the need for a strong theoretical framework, so, too, is it paramount to properly characterize the *behavior* of biological systems. To recapitulate a point made at the beginning, for DNNs to model biological systems it is important to emulate biologically realistic behavior. This does not seem to be the case with CNNs, as the stimulus-response arc is biologically untenable (Edelman 2016).

In this section I have selected examples from Lake et. al. (2017) and Edelman's (2016) work to argue that there are some difficulties in relating (at least) some DNNs to biological systems at the behavioral-level. This thereby casts doubt on the mapping between the two in terms of their *phenomenal descriptions*. If we want DNN models to be explanatory of biological systems, however, then it is important to satisfy a similarity relationship here (ibid.). It is key, then, that DNNs work towards approximating the dynamic and agentic features of biological agents (Marblestone, Wayne, Kording 2016; Baldassarre et. al. 2017). I have tried to argue that there is important work to be done at the behavioral-level in terms of carefully characterizing the behavior of biological systems, and this ultimately goes beyond input-output characterizations (Shagrir, Bechtel 2017). Accordingly, achieving

human-level performance does not seem to establish the required similarity relationship. In the next section, we proceed to the implementational level to assess the mapping relation between biological and artificial neural networks in terms of their *mechanistic descriptions*.

4. The Mechanistic Descriptions: An Analysis at the Implementational-Level

In the previous section, I presented an argument put forth in Lake et. al. (2017). The argument presented herein—targeting the explanatory status of DNN models—overlaps closely with theirs. One important difference, however, is that my argument addresses biological realism and approaches the problem from within the mechanistic framework. As explicated in section (2), the mechanistic framework aims to show how a phenomena of interest is situated in the causal structure of the world. This is done by elucidating the causally relevant entities and activities that give rise to it. When designing a mechanistic model, then, it is important to ensure it reflects not only the behavior of the system, but also its components (the 3M relationship). While Lake et. al. recognize the importance of neuroscientific knowledge in 'constraining' cognitive theories, they argue that a bottom-up approach typically fails to address how structure (synapses and morphology) relates to function (higher-level behavior and cognition). They therefore stress the importance of a strong conceptual framework that constrains future research (2017: 21). The previous section motivated similar concerns.

Nevertheless, I would argue that attention to biological veridicality does more than simply constrain model selection, and is rather integral to the mechanistic framework, as it is ultimately the mechanistic model that explains. Thus, in this section I address DNNs in terms of their implementational details (the mechanistic description). In particular, I address the neural codes involved in biological and artificial neural networks, and identify a potential disjunct between the two (therefore not satisfying the 3M requirement). While there are a variety of codes under investigation, the *rate* and *temporal-codes* dominate the debate (Brette 2015).

Before proceeding, it is important to briefly comment on the role of abstraction in modeling biological systems. Following Godfrey-Smith (2006), I take abstraction as the (un)intentional omission of details that might be causally relevant to the system's functioning, but do not need to be explicitly represented in the model. Abstraction is necessary feature of all computational models, as it makes them both more intelligible and tractable (Kriegeskorte 2015). As such, one may wonder what significance the neural code has for the question at hand. Why neural codes? Why not architectural features, or the fact that neurons are represented as dimensionless points? After all, an analysis of

DNNs and neural systems is in many ways a study in contrasts (Cox, Dean 2014). In principle, then, any number of features could have been identified.

The reason the neural coding debate is a useful example is that it presents intriguing similarities to the argument presented in the previous section. Therein, we saw that characterizing behavior in terms of input-output relations might obscure many features we view as integral to biological agents. Similarly, it has recently been argued that the coding debate revolves around a similar complication (Brette 2015), as I try to make clear below. Moreover, empirical evidence is mounting that suggests spiking dynamics might not only be computationally significant, but necessary for explaining neural systems (Brette 2015). Approximating spiking dynamics in DNNs would not only be more biologically realistic, then, but is "arguable the only viable option if one wants to understand how the brain computes" (Tavanaei et. al. 2018: 1). Thus, while spiking models are not as powerful (in terms of performance) as rate-coding ones, there are reasons to believe this feature is important to capture (ibid.; Huh, Sejnowski 2017; Abbott et. al. 2016)[6].

4.1 Biological and Artificial Neural Networks: A Tale of Two Codes

The neural code is postulated as the means by which neurons encode information in the form of incoming spike trains, and decipher that information in order to emit (or not) an action potential—one of the main forms of neural communication (Dayan, Abbott 2005). At a more abstract level, the neural code concerns how an animal is able to transform information on the sensory periphery into complex, adaptive behavior. As mentioned above, the two most common codes to be discussed are the *rate* and *temporal-codes* (ibid.).

A *rate-code* is an abstract mathematical description of a neuron's electrical activity (usually in the form of differential equations). It is called a rate-code because it captures the neuron's mean firing rate, which is obtained by averaging over many spikes (Brette 2015). Moreover, rate-codes typically operate according to Poisson statistical assumptions, which assumes that the arrival of an action potential is independent of previous input or activity. Under a rate-code, then, the main source of information transmission is the neuron's mean firing rate. Importantly, as we will see below, neurons are intrinsically variable entities, meaning they emit action potentials that often vary drastically from the mean. But, because the average is what is most important to the rate-code, this variability is

[6] It is also important to note that from a machine-learning perspective focused on performance and results, this argument does not apply necessarily. However, when bringing these models to bear on biological systems, mechanistic considerations should apply.

abstracted as 'noise' that plays no causal role in the system. In contrast, a *temporal-code* posits spikes and information processing as *history-dependent* (a non-Poisson feature (Averbeck 2009)). In this case, it is the precisely timed and temporally coordinated arrival and emission of spikes that constitutes the fundament of information processing. Neuronal variability, on this account, is functionally significant (ibid.).

The question in the debate is typically put as follows: in order to understand neural information processing, is it necessary to focus on the individual instantiations of action potentials; or can we essentially approximate spiking dynamics using the firing-rate description? That is, does the firing-rate capture the most important features of spiking dynamics? Brette has recently argued that the question as thus posed might be misleading, however, since it approaches the topic solely from what he calls the 'coding perspective', a view which casts "the question exclusively in terms of… the relationship between stimuli and particular observables (spike trains or rates)" (2015: 2)[7]. Brette argues that such an approach misses an important problem, which is to know which one, rates or spikes, has a *causal role* in the system (ibid.: 8).

In this sense, Brette raises a worry that parallels one in section (3). It will be recalled that characterizing behavior in terms of input-output relations is not sufficient to capture behavior in the awake and situated animal. The justification for this was that couching the organism in terms of input-output relations elided key features of biological systems (Krakauer et. al.'s 'easy mistake'). Similarly, looking at the problem from a 'coding perspective', that is, in terms of inputs and measurable observables, seems to gloss over a key explanatory question: which one, rates or spikes, plays a causal role in the system—which one is relevant to its functioning?

Note how this is a subtly different question from the previous one, as it does not address which description, in terms of inputs and outputs, is best for understanding the system. Instead, it focuses more explicitly on the causal role of computational vehicles. This interlocks nicely with the mechanistic contention that identifying the causally relevant features that realize the explanandum is necessary to *explain* the system. While it is still an open question how exactly the brain encodes information, evidence has been accumulating recently that stresses the importance of history-dependent spikes (Deneve 2008; Izhikevich 2006). Interestingly, current DNNs rely primarily on a

[7] Note the flavor of an input-output doctrine evidenced in the use of 'stimuli' and 'particular observables'.

rate-code (Kriegeskorte 2015) — suggestive perhaps of a mechanistic disjunct between biological and artificial neural networks.

4.2 Can Rate-Code DNNs Approximate Spiking Dynamics?

While including spiking dynamics into DNNs is a burgeoning research topic (Gestner et. al. 2014; Thalmeier et. al. 2015), current DNNs predominantly rely on a rate-code. From a mechanistic perspective, then, it is particularly important to asses whether the abstract mathematical description of a neuron's firing rate reliably approximates spiking dynamics (establishing a 3M relationship). That is, if neuron's *do* rely on non-Poisson processes then we must ask whether rate-coding DNNs reliably approximate these features. As I now try to suggest, it might be difficult to establish such a mapping relationship.

One contentious feature of the rate-code is that it relies on Poisson statistical assumptions, briefly sketched above. This entails that action potentials and spike trains are fundamentally history-independent, with the average firing-rate constituting the main source of information. Accordingly, the intrinsic stochasticity or spontaneity that characterizes nervous activity is abstracted out in a rate-code DNN and instead conceptualized as neuronal noise. If anything, noise compromises a neuron's ability to communicate effectively, and is thus seen as inconsequential to neural communication (Brette 2015). Recently, however, it has been argued that the non-Poisson features associated with spiking dynamics (i.e., the history-dependent nature of neural activity and computational significance of precisely timed and variable spikes) are not only causally relevant, but integral to our understanding of biological systems (Averbeck 2009). In this sense, the variability of a neuron's firing-rate (noise according to the former view) is actually important in terms of information transmitted. As an example, under Debanne et. al.'s predictive coding account, which focuses on spiking dynamics, noise codes for uncertainty (2013). Therefore, abstracting this feature out with a rate-code would miss an important feature of the target system.

An important upshot of this is that in order for a rate-code to reliably approximate spiking dynamics, one would have to establish a homogeneity between the actual spiking behavior of individual neurons embedded in biological neural networks and the abstracted rate-code in DNN models, which operates over many layers and units (called a continuous activation value (Tavanaei et. al. 2018)). In other words, the intrinsically stochastic nature of individual neurons is homogenized under a rate-code, as variability is factored out as noise (Brette 2105). This does not seem to reflect the staggering heterogeneity of neural activity in the brain. Neurons are, after all, Heraclitean in nature

(Chirimuuta 2018), constantly shaped by history-dependent activity due to an organism's interaction with the environment. Thus, if what is important for adaptive and flexible behavior is in fact the non-Poisson features of history-dependent spiking dynamics, the rate-code might not approximate this. After all, history-dependence "is a fundamentally non-Poisson feature, as the spike times in a Poisson process would not depend on the history of prior spikes" (Averbeck 2009: 310).

As a final note, it is interesting to look at the experimental setups in which many early rate-coding studies were performed. Often times, these involved recording *in vitro* using neurons in a petri dish, subjecting animals to constrained environments where few choices were available to them[8], or on the anesthetized animal (Brette 2015). Such experiments often evidenced a stereotypy to neuronal activity that has not transferred over to experimental setups in which animals openly navigate their environment. That is, as technology advanced and allowed researchers to experiment on awake and active animal, studies began to appear suggesting the importance of precisely timed spikes (ibid.). This is largely due to the fact that the timescales at which both processes occur vary quite dramatically. For example, studies on sensory and motor areas, such as auditory cortex and parietal lobe (Averbeck 2009; Maimon & Assad 2009), show that neurons interact on short, millisecond timescales. Conversely, the processes required to average over many spikes exceeds this and occurs, instead, in seconds (a rather long time in neuroscientific terms). This is now generally seen as an unlikely option to account for quick, robust animal behavior (Brette 2015). Thus, as the experimental setups became more realistic (in terms of an animal's mobility and behavioral assays), evidence began to appear suggesting many non-Poisson features of neural activity (ibid.).

While more could be said on the rate vs. temporal-coding debate (cf. esp. Brette 2015; Softky, Koch 1993; Dayan, Abbott 2005), the above motivates just a few reasons for thinking that a rate-coding model might not map onto the causal features of spiking biological neural networks. This is namely because spiking dynamics, which causally relies on the arrival of temporally coordinated and history-dependent spikes, is abstracted out under the rate-code. Insofar as a 3M relationship requires a mapping between features of the model (in this case, a continuous activation value or rate-code (Tavanaei et. al. 2018)) and the entities or activities of the target system (i.e., spiking dynamics), then

[8] Interestingly, constraining the environment of an organism characterizes the experiments used to corroborate the stimulus-response arc. The problem with these early experiments was precisely the fact that they dealt with toy-environments that were easy to control (for technological reasons, experiments in behaving animals was too difficult), but did not scale up to the real world dynamics of animal's activity (Godfrey-Smith 2016).

it is important to identify whether or not the former approximates the latter. I have tried to suggest a few difficulties in doing so herein.

Ultimately, a good model should be like a good caricature: it emphasizes salient features while underplaying others (Abbott 2008). Accordingly, a satisfactory mechanistic model does not require impeccable biological accuracy to qualify as explanatory. Indeed, it is often the case that a stringent focus on such accuracy might compromise intelligibility, making these models 'the least satisfying' (O'Leary et. al. 2015: 87). Similar concerns motivate Kriegeskorte (2015) to note that

> Merely pointing out a difference to biological brains… does not constitute a legitimate challenge [to DNNs].… [The] fact that real neurons spike does not pose a challenge to a rate-code model…. [However,] if spiking were a computational requirement, and spiking models outperformed the best rate-code model… then this model would present a challenge to the rate-code approach. (2015: 438)

Again, it should be made clear that I am not rejecting that abstraction is a major feature of model building (Bechtel, Levy 2013). Nevertheless, from a mechanistic perspective, if we want our DNN models to explain salient features of biological systems, then that seems to require representing features of the target system that play a *causal role* in its realization. In this section, I have tried to motivate reasons for thinking spikes are in fact computationally important, but nevertheless omitted from the rate-coding approach that currently predominates in machine-learning. From a mechanistic perspective, what we are concerned with in understanding neural circuits is what they are *actually* doing, and in this sense it is important to determine a mapping relationship between the two systems (Craver 2007). The current relationship between DNNs and brains might not therefore be appropriately expressed in the approximation of a rate-code, and a mapping relationship is accordingly called into question.

5. Conclusion: Deep Neural Networks as How-Possibly Models

To conclude, the goal of this paper has been manifold. First and foremost, I set out to provide an explanatory framework in which to assess status of explanations provided by current generation deep neural networks. The framework I assumed was the *mechanistic philosophy of science*. According to the mechanistic framework, a model has to map onto its target system at (at least) two points in order to qualify as explanatory: the *behavioral* and *implementational* levels. These were called the *phenomenal* and *mechanistic* descriptions, respectively. With this as my starting point, I began to sequentially address each level, ultimately suggesting difficulties for a mapping relationship between the two systems. As

such, DNNs, as they stand, would qualify as *how-possibly* models with 'explanatory purport', but only loosely constrained by behavioral and biophysical considerations (sections 3 & 4; Craver 2007).

This should be an unsurprising conclusion. After all, nobody expects first-generation DNNs to reliably approximate biological neural networks at the start: "it would be hasty to judge the merits of DNNs based on the level of abstraction chosen in the first-generation. The usage of DNNs in computational neuroscience is still in its infancy. Integration of biological details will require close collaboration between modelers, experimental neuroscientist, and anatomists" (Kietzmann et. al. 2018). Thus, while one of my goals was to assess the current status of DNNs as models of behavior and neural information processing (requiring an assessment of the current level of abstraction in DNNs), I would like to propose a positive point to conclude. As I mentioned at the beginning, DNNs *can* contribute to the explanatory endeavors of neuroscientific research. In order to do so, however, it is important that they approximate key features of biological systems (van Gerven 2017: 14).

For this reason, I suggested two areas that are relevant for understanding and explaining animal behavior and neural communication: characterizing behavior in terms of an animal's learning tasks (Lake et. al. 2017) and its active, self-initiated nature (Marblestone, Wayne, Kording 2016; Baldassarre 2011; Edelman 2016); and the importance of approximating spiking dynamics, as the temporally coordinated arrival and spontaneous elicitations of individual action potentials is increasingly seen as computationally significant (Brette 2015). This in turn requires a careful and judicious analysis of the many things animals *do*, as well as identifying the features that are *relevant* to them.

At both levels we require a proper specification of the organism and its activities, which will ultimately revolve around identifying what is *relevant* to that system itself, whether it is behavior or the causal role entities and activities play in its instantiation. This in turn requires a strong conceptual framework that effectively and reliably guides future research into mechanistic underpinnings. Ultimately, in using DNNs as explanatory proxies *of* biological systems, questions concerning biological realism should always be present. This necessitates not only biologically realistic parts and activities, but realistic behavior, both of which go beyond input-output characterizations, as I tried to defend above. As DNNs become more biologically realistic at both levels, then, they promise to assist in describing, understanding, and *explaining* biological systems.

References

Abbott, L.F. 2008. Theoretical Neuroscience Rising. *Neuron*, vol. 60: 489-495.

Abbott, L.F., Depasquale, B., Memmesheimer, R-M. 2016. Building functional networks of spiking model neurons. *Nature Neuroscience*, vol. 19. 350-355.

Anderson, D.J. & Perona, P. 2014. Toward a science of computational ethology. *Neuron*, vol. 84: 18-31.

Averbeck, B.B. 2009. Poisson or Not Poisson: Differences in Spike Train Statistics between Parietal Cortical Areas. *Neuron*, vol. 62: 310-311.

Baldassarre, G. 2011. What are Intrinsic motivations? A Bioloigcal Perspective. In: Proceedings of the International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob-2011), ed. A. Cangelosi, J. Triesch, I. Fase

Baldassarre, G., Santucci, V., Cartoni, E., Caligiore, D. 2017. The architecture challenge: Future artificial intelligence systems will require sophisticated architectures, and knowledge of the brain might guide their construction. Response to Lake et. al. "Building machines that learn and think like people". *Brain and Behavioral Sciences*, e253: 25-26.

Barak, O. 2017. Recurrent neural networks as versatile tools of neuroscience research. *Current Opinion in Neurobiology*, vol. 46: 1-6.

Bassett, D., Zurn, P. Gold, J. 2018. On the nature of models in network neuroscience. *Nature Reviews Neuroscience*: https://doi.org/10.1038/s41583-018-0038-8

Bechtel, W. 2008. *Mental Mechanisms: Philosophical perspectives on cognitive neuroscience.* New York, NY, US: Routledge/Taylor & Francis Group

Bechtel, W. 2009. Looking down, around, and up: Mechanistic explantion in psychology. *Philosophical Psychology*, vol. 22: 543-564.

Bechtel, W. & Levy, A. 2013. Abstraction and the Organization of Mechanisms. *Philosophy of Science*, vol. 80: 241-261.

Bermudez, J.L. 2005. *Philosophy of Psychology: A Contemporary Introduction.* New York: Routledge.

Bornstein, A. 2016. Is Artificial Intelligence Permanently Inscrutable? Nautil.us, issue 40.

Bowers, K.S. 1973. Situationism in psychology: An analysis and a critique. *Psychological Review*, vol. 80: 307-336.

Brette, R. 2015. Philosophy of the Spike: Rate-Based vs. Spike-Based Theories of the Brain. *Frontiers in Systems Neuroscience*, vol. 9: 1-14.

Carlson, T. Goddard, E., Kaplan, D., Klein, C., Ritchie, B. 2017. Ghosts in machine learning for

cognitive neuroscience: Moving from data to theory. *Neuroimage*, vol. 30: 1-13/

Chirimuuta, M. 2014. Minimal models and canonical neural computations: the distinctness of computational explanation in neuroscience. *Synthese*, vol. 191: 127-153.

Chomsky, N. 1959. Review of Verbal Behavior. *Language*, vol. 35: 26–58.

Churchland, P., Sejnowski, T. 2016. Blending Computational and Experimental Neuroscience. *Nature Reviews*, vol. 17: 667-668.

Ciresan, D., Meier, U. & Schmidhuber, J. (2012) Multi-column deep neural networks for image classification. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, June 16–21, 2012, pp. 3642– 49. IEEE. [aBML].

Cooper, R.P. & Peebles, D. 2015. Beyond single-level accounts: the role of cognitive architectures in cognitive scientific explanation. *Topics in Cognitive Science*, vol. 7: 243-258.

Cox, D.D., Dean, T. 2014. Neural Networks and Neuroscience-Inspired Computer Vision. *Current Biology*, vol. 24: R921-R929.

Craver, C. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. New York: Oxford University Press.

Craver, C., Kaplan, D., 2018. Are More Details Better? On the Norms of Completeness for Mechanistic Explanation. *The British Journal for the Philosophy of Science*: https://doi.org/10.1093/bjps/axy015.

Crick, F. 1989. The recent excitement about neural networks. *Nature* vol. 337: 129-132 [aBML0]

Dayan, P. & Abbott, L.F. 2005. *Theoretical Neuroscience*. Cambridge, Massachusetts: MIT Press.

Debanne, D., Bialowas, A, Rama, S. 2013. What are the mechanisms for analogue and digital signaling in the brain? *Nature Review Neuroscience*, vol. 14: 63-69.

Deneve, S. 2008. Bayesian spiking neurons I: inference. *Neural computation*, vol. 20: 91-117.

Dong, Q., Wang, H., Hu, Z. 2018. Commentary: Using goal-driven deep learning models to understand sensory cortex. *Frontiers in Computational Neuroscience*, vol. 12:4 1-2.

Doshi-Velez, F. & Kim, B. 2017. Towards a Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608v2 [stat.ML].

Edelman, S. 2016. The minority report: some common assumptions to reconsider in the modelling of the brain and behaviour. *Journal of Experimental & Theoretical Artificial Intelligence*, 28: 4, 751-776.

Fetsch, C.R. 2016. The importance of task design and behavioral control for understanding the neural basis of cognitive functions. *Current Opinion in Neurobiology*, vol. 37: 16-22.

Frank, M.J., & Badre, D. 2015. How cognitive theory guides neuroscience. *Cognition*, vol. 135: 14-20.

Fregnac, Y. 2017. Big-data and the industrialization of neuroscience: A safe roadmap for understanding the brain? *Science*, vol. 358: 470-477.

Gershman, S., Horvitz, E., Tenenbaum, J. 2015. Computational rationality: a converging paradigm for intelligence in brains, minds, and machines. *Science*, vol. 349: 273-278.

Gerstner, W., Kistler, W.M., Naud, R., Paninski, L. 2014. *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge: Cambridge University Press.

Glaser, J., Benjamin, A., Farhoodi, R., Kording, K. 2018. The Roles of Supervised Machine Learning in Systems Neuroscience. arXiv:1805.08239.

Glennan, S. 1996. Mechanisms and the Nature of Causation. *Erkenntnis*, vol. 44: 49-71.

Glennan, S. 2005. Modeling Mechanisms. *Studies in History and Philosophy of Biological and Biomedical Science*, vol. 36: 443-464.

Glennan, S. 2017. *The New Mechanical Philosophy*. Oxford University Press, Oxford.

Godfrey-Smith, P. 2006. The strategy of model-based science. *Biology and Philosophy*, vol. 21: 725-740.

Gomez-Marin, A., Paton, J., Kampff, A., M Costa, R., Mainen, Z. 2014. Big behavioral data: psychology, ethology and the foundations of neuroscience. *Nature Neuroscience*, vol. 17: 1455-1461.

Graham, G. 2017. Behaviorism. In *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), Edward N. Zalta (ed.).

Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. 2017. Neuroscience-Inspired Artificial Intelligence. *Neuron Review*, vol. 95: 245-258.

Heisenberg, M. 2014. The beauty of the network in the brain and the origin of the mind in control of behavior. *Journal of Neurogenetics*, vol. 28: 389-399.

Helmstaedter, M. 2015. The Mutual Inspiration of Machine-Learning and Neuroscience. *Neuron*, vol. 86: P25-P28.

Herz, A., Gollisch, T., Machens, C., Jaeger, D. 2006. Modeling Single-Neuron Dynamics and Computations: A Balance of Detail and Abstraction. *Science*, vol. 314: 80-85.

Hinton, G. 2007. Learning multiple layers of representation. *TRENDS in Cognitive Science*, vol. 14: 428-434.

Huh, D. & Sejnowski, T. 2017. Gradient descent for spiking neural networks. Arxiv:1706.04698: 1-10.

Izhikevich, E.M. 2006. Polychronization: computation with spikes. *Neural Computation*, vol. 18: 245-282.

Jazayeri, M. & Afraz, A. 2017. Navigating the Neural Space in Search of the Neural Code. *Neuron*, vol. 93: 1003-1014.

Katz, B. & Miledi, R. 1967. The Timing of Calcium Action During Neuromuscular Transmission. *Journal of Physiology*, vol. 189: 535-544.

Kaplan, D. 2011. Explanation and description in computational neuroscience. *Synthese*, vol. 183: 339-373.

Kaplan, D. & Craver, C. 2011. The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective. *Philosophy of Science*, vol. 78: 601-627.

Kay, K. 2017. Principles for models of neural information processing. *Neuroimage*, vol. 30: 1-9.

Khaligh-Razavi, S., Kriegeskorte, N. 2014. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, vol. 10: 1-29.

Kietzmann, T., McClure, P., Kriegeskorte, N. 2017. Deep Neural Networks in Computational Neuroscience. *bioRxiv*: https://doi.org/10.1101/133504

Krakauer, J., Ghazanfar, A., Gomez-Marin, A., MacIver, M., Poeppel, D. 2017. Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*, vol. 93: 480-490.

Kriegeskorte, N. 2015. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, 2015, 1: 417-446.

Kubilius, J., Bracci, S., Op de Beeck, H. 2016. Deep Neural Networks as a Computational Model for Human Shape Sensitivity. *PLoS Computational Biology*: doi:0.1371/journal.pcbi.1004896

Lake, B., Ullman, T., Tenenbaum, J., Gershman, S. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, e253: 1-72.

LeCun, Y., Bengio, Y., Hinton, G. 2015. Deep Learning. *Nature*, vol. 521: 436-443.

Lehky, S. & Tanaka, K. 2016. Neural representations for object recognition in inferotemporal cortex. *Current Opinion in Neurobiology*, vol. 37: 23-35.

Machamer, P., Darden, L., Craver, C. 2000. Thinking about Mechanisms. *Philosophy of Science*, vol. 57: 1-25.

Maimon, G. & Assad, J. 2009. Beyond Poisson: Increased Spike-Time Regularity across Primate Parietal Cortex. *Neuron*, vol. 62: 426:440.

Majaj, N. & Pelli, D. 2018. Deep learning: using machine learning to study biological vision. *bioRxiv*: doi:http://dx.doi.org/10.1101/178152.

Marblestone, A., Wayne, G., & Kording, K. 2016. Towards an Integration of Deep Learning and Neuroscience. *Frontiers in Computational Neuroscience*, vol. 10: 1-41.

Marcus, G. 2018. Deep Learning: A Critical Appraisal. arXiv:1801.00631v1 [cs.AI].

Matthewson, J. 2017a. Models of Mechanisms. In Glennan, S., & Illari, P. *The Routledge Handbook for Mechanisms and Mechanistic Philosophy*: 225-235.

Matthewson, J. 2018. Detail and generality in mechanistic explanation. *Studies in History and Philosophy of Science*, vol. 30: 1-9.

Milkowski, M. 2013. *Explaining the Computational Mind*. MIT Press. Cambridge, Massachusetts.

Milkowski, M. 2016. Explanatory completeness and idealization in large brain simulations. *Synthese*, vol. 193: 1457-1478.

O'Leary, T., Sutton, A., & Marder, E. 2015. Computational models in the age of large datasets. *Current Opinion in Neurobiology*, vol. 32: 87-94.

Piccinini, G. & Craver, C. 2012. Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese*, vol. 183: 283-311.

Ross, L.N. 2015. Dynamical Models and Explanations in Neuroscience. *Philosophy of Science*, vol. 67: 1-25.

Rumelhart, D. & McClelland, J. 1987. *Parallel Distributed Processing Volume 1*. MIT Press, Cambridge Massachusetts.

Salmon, W. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Sak, H., Senior, A., Beaufays, F. 2014. Long short-term memory based recurrent neural networks architectures for large vocabulary speech recognition. arXiv:1402.1128 [cs.NE].

Sejnowski, T., Churchland, P., & Movshon, J.A., 2014. Putting big data to good use in neuroscience. *Nature Neuroscience*, vol. 13: 1440-1441.

Silberstein, M. & Chemero, A. 2013. Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences. *Philosophy of Science*, vol. 80: 958-970.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G. V. D., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K, Graepel, T. & Hassabis, D. (2016) Mastering the game of go with deep neural networks and tree search. Nature 529(7585):484–89. [arBML, MB]

Shagrir, O. & Bechtel, W. 2017. Marr's Computational Level and Delineating Phenomena. In Kaplan, D. *Explanation and Integration in the Mind and Brain Sciences*: 190-214. Oxford University Press.

Skinner, B.F. 1957. *Verbal Behavior*. Copely Publishing Group.

Softky, W. & Koch, C. 1993. The highly irregular firing of cortical cells is inconsistent with temporal integration of random ESPSs. *Journal of Neuroscience*, vol. 13: 334-350.

Sompolinsky, H. 2014. Computational neuroscience: beyond the local circuit. *Current Opinion in Neurobiology*, vol. 25: xiii-xviii.

Sudhof, T. 2000. The Synaptic Vesicle Cycle Revisited. *Neuron*, vol. 28: 317-320.

Sutskever, I., Vinyals, O., QVV, Le. 2014 Sequence to sequence learning with neural networks. *Advanced Neural Information Processing Systems*, vol. 27:3 104-112.

Tavanaei, A., Ghodrati, M., Kheradpisheh, S.R., Masquelier, T., Maida, A. 2018. Deep learning in spiking neural networks. arXiv:1804.08150v2.

Thalmeier, D., Uhlmann, M. Kappen, H., Memmesheimer, R.M, May, N.C. 2015. Learning universal computations with spikes. ArXiv:1505.07866v1: 1-35.

Thompson, E. 2007. *Life in Mind: Biology, Phenomenology, and the Sciences of Mind*. Harvard: Belknap Harvard University Press.

van Gerven, M. 2017. Computational Foundations of Natural Intelligence. *Frontiers in Computational Neuroscience*, vol. 11: 1-23.

Woodward, J. 2017. Explanation in Neurobiology: An Interventionist Perspective. In Kaplan, D. *Explanation and Integration in Mind and Brain Sciences*. New York: Oxford University Press.

Woodward, J. 2003. *Making Things Happen*. New York: Oxford University Press.

Yamins, D., Hong, H., Cadieu, C., Solomon, E., Seibert, D., Dicarlo, J. 2014a. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Science*, vol., 111: 8619-8624

Yamins, D. & DiCarlo, J. 2016a. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, vol. 19: 356-364.

Yamins, D. & DiCarlo, J. 2016b. Eight open questions in the computational modeling of higher sensory cortex. *Current Opinion in Neurobiology*, vol. 37: 114-120.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H. 2015. Understanding neural networks through deep visualization. arXiv:1506.06579.

Zednik, C. 2014. Are Systems Neuroscience Explanations Mechanistic? Paper presented at *Philosophy of Science Association 24th Bieenial Meeting (Chicago, IL),* November 2014.

Zednik, C. 2015. Heuristics, Description, and the Scope of Mechanistic Explanation. In C. Malaterre & P-A. Braillard (Eds.), *Explanation in Biology: An Enquiry into the Diversity of Explanatory Patterns in the Life Sciences* (pp. 295-318). Dordrecht: Springer.

Zednik, C. 2018. Computational Cognitive Neuroscience. In M. Colombo & M. Sprevak (Eds.), *The Routledge Handbook of the Computational Mind*. London: Routledge.