Prediction Versus Understanding in Computationally Enhanced Neuroscience

M. Chirimuuta (forthcoming) *Synthese*

ABSTRACT

The use of machine learning instead of traditional models in neuroscience raises significant questions about the epistemic benefits of the newer methods. I draw on the literature on model *intelligibility* in the philosophy of science to offer some benchmarks for the *interpretability* of artificial neural networks (ANN's) used as a predictive tool in neuroscience. Following two case studies on the use of ANN's to model motor cortex and the visual system, I argue that the benefit of providing the scientist with understanding of the brain trades off against the predictive accuracy of the models. This trade-off between prediction and understanding is better explained by a non-factivist account of scientific understanding.

1. INTERPRETABILITY AND INTELLIGIBILITY IN BIG-DATA NEUROSCIENCE

Neuroscience is undergoing a big-data revolution, where high throughput methods generate terabytes of neural recordings, and machine learning algorithms are at work searching for meaning and pattern amongst the endless numbers of simultaneously recorded spikes and traces. Responses to these innovations have been both enthusiastic and tepid (Churchland and Sejnowski 2016; Paninski and Cunningham 2018; Fre□gnac 2017: 471). This paper attempts to deliver a non-partisan analysis of the advantages and limitations of certain applications of machine learning in neuroscience, looking in particular at artificial intelligence based on connectionist networks (*artificial neural networks* or ANN's)[1] used to model the responses of neurons in visual and motor areas. I will argue that while the predictive accuracy of such models is in a different league from that of previous generations of hand-coded models, this comes at a cost of the understanding of the neural systems afforded by modelling them. In other words, there is a trade-off between a model's predictive power and its ability to increase the scientist's understanding of a neural response.

[1] See Buckner (2019) for an accessible review of deep learning.

Philosophers such as Carl Hempel took understanding to be subjective and peripheral to philosophy of science,[2] but it is common for scientists themselves to characterise understanding as central to their endeavour. One example is the pioneering neuroscientist, Emil du Bois-Reymond (1874) who argued, not without controversy, that the limits of our capacity to understand nature are the limits of science itself. In their statement of aims for the US government funded BRAIN Initiative, Jorgenson and co-authors write that understanding is the ultimate goal of neuroscientific research:

> The overarching goal of theory, modelling and statistics in neuroscience is to create an *understanding* of how the brain works—how information is encoded and processed by the dynamic activity of specific neural circuits, and how neural coding and processing lead to perception, emotion, cognition and behaviour. [Emphasis added. Quoted in Fairhall and Machens (2017: A1)]

The topic of understanding has recently risen in importance within the philosophy of science.[3] This literature is helpful not only for analysis of the goals of neuroscience, but also in current debates on the interpretability of AI, and other complex computational tools, for human users. I will argue that philosophical accounts of understanding (*of natural systems*) and intelligibility (*of theories or models*) help to shed light on the current discussion, within computer science, over model interpretability – the question of how to make the decisions and classifications generated by AI comprehensible to human users.

Given the trade-off, to be presented in the next section, between prediction and understanding afforded by computational models in neuroscience, I will argue that a non-factive account of understanding best suits the case in hand. Roughly speaking, non-factivists about understanding do not equate understanding with the learning of facts about nature, or the knowledge of true explanations; rather, scientific understanding is a matter of skill (de Regt 2017) or an epistemic benefit that is more often than not conferred by idealisations rather than literally true representations of nature (Potochnik 2017; Elgin 2017).

[2] See Hempel (1965: 413) discussed in De Regt (2017: 16). See also Hooker and Hooker (2018) on scientific realism and the requirement that science produce interpretable models that go beyond "naked prediction."

[3] Four books recently published are Khalifa (2017), De Regt (2017), Elgin (2017) and Potochnik (2017). De Regt and Potochnik defend the view that understanding is the central epistemic aim of science.

1.1 Interpretability as Intelligibility

Computer scientist Zachery Lipton (2016) observes that while there is a consensus that model interpretability is a good thing, there is no convergence on one definition or operationalization. Most discussions focus on the ability of non-expert users to see the reasons behind an AI's decisions. I put this issue aside as I am concerned with the capacity of expert users, including the model builders themselves, to know about the processes taking place between input and output of a complex, trained neural network. Amongst the many facets of interpretability discussed by Lipton, the one relevant to my study is the notion of interpretability as *transparency*, which he calls "the opposite of *opacity* or *blackbox-ness*" (2016:4). "Black box"[4] is a common, if colloquial term, for a device or piece of code which transforms inputs into outputs without providing any indication of the method behind this operation. The black box flavour of artificial neural networks is something discussed by experts within neuroscience. For example Omri Barak (2017:5) points out that "machine learning provides us with ever increasing levels of performance, accompanied by a parallel rise in opaqueness".

However, it would be wrong to say that ANN's are literally black boxes because so many features of their internal architecture and workings are known to the model builders. At the same time, the exact way that a network arrives at predictions or classifications is often quite opaque to its makers, hence the concerns. As theoretical neuroscientists Gao and Ganguli (2015:151) describe matters,

> Each of these [artificial neural] networks can solve a complex computational problem. Moreover, we know the full network connectivity, the dynamics of every single neuron, the plasticity rule used to train the network, and indeed the entire developmental experience of the network….. Yet a meaningful understanding of how these networks work still eludes us, as well as what a suitable benchmark for such understanding would be.

The issue is how to characterise the relative degrees of transparency and opacity exhibited by different models, and to explain the specific benefits of more

[4] One might worry that the term "black box" is most often used pejoratively, to dismiss an algorithm or model that the speaker happens not to understand because she lacks expertise. (I thank Michael Tarr for this point.) I emphasize here that I am considering what is comprehensible to a maximally well-informed human, and that I do not take "black boxes" to be bad by definition. They certainly have their uses in science and engineering.

transparent models. I propose that the notion of *intelligibility* of scientific theories advanced by Henk de Regt is a helpful starting point.

De Regt and Dieks (2005:143) make the point that a perfect black box predictor of empirical observations – an oracle – would not count as a scientific theory because it lacks intelligibility:

> In contrast to an oracle, a scientific theory should be intelligible: *we want to be able to grasp how the predictions are generated, and to develop a feeling for the consequences the theory has in concrete situation.* (emphasis original)

On this account, intelligible theories enable scientists to build models which explain natural phenomena and thereby yield understanding (de Regt 2014:32). Thus intelligibility is not a mere psychological add-on, but is fundamental to scientists' ability to use theories.

When characterising intelligibility, de Regt employs a notion introduced by Werner Heisenberg and endorsed by Richard Feynman:

> Criterion of Intelligibility (CIT): A scientific theory *T* (in one or more of its representations) is intelligible for scientists (in context *C*) if they can recognize qualitatively characteristic consequences of *T* without performing exact calculations. (de Regt 2014:33; also De Regt and Dieks (2005: 151ff)

Gao and Ganguli also make the point that accurate prediction does not entail understanding, and refer to the same criterion introduced in physics:

> we understand a physical theory if we can say something about the solutions to the underlying equations of the theory without actually solving those equations. (Gao and Ganguli 2015: 148)

The idea is that to achieve comprehension, going beyond the bare ability to make accurate predictions, the scientist must have a reliable sense of under what circumstances the classes of predicted phenomena obtain, even before running the calculations. The qualitative description of the system provided by an intelligible theory goes beyond the bare relationships between input and output variables that are supplied by a black-box.

For example, a simple, basically linear model of a visual neuron's receptive field yields the qualitative expectation, prior to any calculation, that an increase in the amount of light falling in the excitatory field of the neuron, will result in an increase in the model's predicted response. In addition, a theory or model that affords visualisation of the target system is typically more intelligible than a purely

abstract theory because other mathematical or concrete models can be constructed on the basis of such intuitive pictures (de Regt 2009:33; de Regt and Dieks 2005:155). As we will see below, intelligible mathematical models of neural responses often suggest simple circuit diagrams that illustrate a combination of excitatory and inhibitory inputs which would bring about the responses described in the model. I will take this to be a kind of visualisability.

It is a feature of de Regt's account that intelligibility is not an intrinsic property of theories but is relative to the scientific context – "the capacities, background knowledge, and background beliefs of the scientists" using the theory (de Regt 2009:33). Thus there is a body of skills and knowledge that a trained scientist can employ in order to render a theory intelligible, and this is not restricted to visualisation. The context-relativity of intelligibility will be important to my discussion in Section 3.2, when I consider whether, with future developments in reverse engineering, ANN's are going to become more intelligible. I should also point out that in physics, where the criterion presented by de Regt has its origin, there is a clear separation between fundamental theory and models. According to De Regt, intelligible theories in physics allow scientists to build models that explain target systems. In the process of theory use and model building, the scientist comes to understand the system. Within computational neuroscience, the terms "theory" and "model" are often used interchangeably. Because this discipline lacks the strict theory-model division of labour seen in physics, I will speak of the quantitative representation of the neural system -- the 'theory/model' -- as providing explanations and thereby understanding.[5]

1.2 Decoding the Brain

My focus is on a tradition of research that builds mathematical models of neurons' response profiles, aiming both at predictive accuracy and at theoretical understanding of the computations performed by classes of neurons. The book *Spikes: Exploring the Neural Code* (Rieke et al. 1999) has served as an important reference point for researchers because it gives the question of what it takes to "understand the neural code" a precise answer – it is the ability to *decode* spike trains, to interpret a string of neural pulses in terms of external conditions

[5] Below in the case studies I write about neuroscientific 'theories' or 'models,' following the scientists' use. Bear in mind that I mean these terms usually to refer to the undifferentiated class, theory/model.

represented by that activity. This decoding problem places the two goals of understanding and prediction of neural responses at the heart of research in theoretical neuroscience, as Stevenson and Kording (2011: 140) describe:

> Understanding what makes neurons fire is a central question in neuroscience and being able to accurately predict neural activity is at the heart of many neural data analysis techniques. These techniques generally ask how information about the external world is encoded in the spiking of neurons. On the other hand, a number of applications, such as brain-machine interfaces, aim to use neural firing to predict behavior or estimate what stimuli are present in the external world. These two issues are together referred to as the neural coding problem. (citations omitted)

Research in this area finds a practical outlet in brain machine interface (BMI) technology, [6] where a prominent application under development is for rehabilitation devices which record activity via microelectrodes implanted in the primary motor cortex (M1),[7] decode the pattern in a computer, and use the decoded signal to control movements of a robotic limb or cursor. Another kind of decoding system, widely discussed under the name of "mind-reading" takes non-invasive fMRI data from the visual cortex or from brain areas involved in semantic processing, in order to reconstruct the visual experience or the subject matter seen/thought by an individual.[8]

Theories/models of what information neurons *encode* about the external world have featured in the design of computer programmes at the heart of such devices. These programmes are known as *decoders*. In the case of motor-BMI's, decoders are algorithms that map neural activity to kinematics of a cursor or robotic limb. An *encoding theory/model* takes the form of a function mapping an external state to a neural response pattern:

(1) *Neural Response = f (external state)*

[6] For technical details see Nicolas-Alonso and Gomez-Gil (2012) and references therein; for philosophical discussion see Datteri (2008) (2017), Craver (2010) and Chirimuuta (2013).

[7] There are BMI's which record from other areas of the motor system; for ease of presentation I refer only to M1 interfaces. Likewise, I focus on motor BMI's reliant on invasive, intra-cortical recordings.

[8] See e.g. Nishimoto et al. (2011) and Naselaris and Kay (2015).

As schematised in Equation 1, the encoding model is a means of predicting how a neuron will respond if presented with a certain stimulus.[9] This raises the question of whether accurate neural prediction helps to advance any further epistemic goals, such as explanation and understanding. In computational neuroscience, functions of the sort referred to in Equation 1 have been thought of as not merely describing the mathematical relationship between sets of variables, *but* as *characterising the computation that is itself performed by a neuron or neuronal population*. On the assumption that the brain is an information processing device, primary explanatory goals have been to work out which computations it carries out, and why (Chirimuuta 2014; 2018).[10] In the case of motor cortex, the question of what kind of information its neurons encode is still a matter of controversy (Omrani et al. 2017), and may be illuminated by building encoding models.

While there are numerous ways to apply machine learning in neuroscience (Glaser et al. 2019), my case studies are restricted to the specific application of ANN's for the development of encoding models and decoders. Thus my finding of a trade-off between prediction and understanding does not immediately generalise to all kinds of modelling within neuroscience, but covers instances of models intended to represent the functions computed by neural systems, and hence are candidates for providing understanding of the neural code. It needs emphasis that the ANN's discussed in my case studies are intended to represent *neural computations* – the encoding functions -- and *not neural anatomy or physiology*. For example, the nodes and connections in the ANN for the motor cortex (Section 2.1) should not be taken to represent, even in a highly simplified manner, a concrete population of biological neurons and connections amongst them. Instead, what the model is intended to duplicate from the neural system is its abstract computational or information processing capacity. Stinson (forthcoming) gives a helpful account of ANN's where the target of representation *is* the material brain. This is an appropriate account for studies such as those from the di Carlo lab at MIT, where the target of representation of the deep neural network is the primate ventral stream, considered physiologically, computationally and anatomically, with its

[9] Equally, one could think of this in terms of retrodiction of a motor state: *what was the activity in the motor cortex neuron that preceded the rightwards movement of the arm?* For convenience I will speak just in terms of prediction.

[10] Elsewhere I criticise this literalist interpretation of neuro-computational models, but for the purpose of this paper it is taken as granted (Chirimuuta forthcoming).

characteristic hierarchical structure (Yamins and DiCarlo 2016).[11] In contrast, my cases utilise ANN's not because of their brain-inspired features, but purely because of their mathematical property of being able to approximate any non-linear function that relates one dataset to another, such as the mapping between visual stimuli and neural responses – their property of being "universal function approximators" (Kriegeskorte 2015: 422-23). As I will argue, this is the source of their predictive power, while the fact that the function discovered by the ANN is embedded in the connection weights of the trained network, and not delivered explicitly to the scientist, means that the ANN lacks what Creel (forthcoming) calls "functional transparency". This renders them more opaque, and less intelligible, than the unsophisticated hand-coded models developed previously by neuroscientists.

Physiological inspection of data generated from electrophysiological recordings of visual and motor neurons under naturalistic stimulation and movement conditions has given scientists good reason to think that the computations performed by those systems are very complex and nonlinear. But given the complexity of the brain, it is surprising that many of the models that have been quite predictively accurate -- albeit in a limited range of non-naturalistic experimental conditions -- are simple linear functions (or linear models with straightforward nonlinear additions) that present no interpretative difficulties.[12] This meant that until recently, in visual and motor neuroscience, there has been a fruitful co-alignment of the goals of prediction and understanding. The story I will tell in the next section is one of an emerging misalignment of those goals: compared to 20 years ago, the most predictively accurate models make less of a contribution to the project of understanding the brain. I will present two cases that illustrate the tendency towards divergence, then argue in Section 3 that there is a trade-off between the predictive accuracy and understanding afforded by the models. Section 4 will relate this finding to the literature on explanation in computational neuroscience and say why a non-factive account of scientific understanding helps to make sense of the trade-off.

[11] But note that this group also develops encoding models of ventral stream neurons. The trade-off between prediction and understanding applies to some aspects of their work.

[12] See Carandini et al. (2005) for examples and discussion.

2. TWO CASE STUDIES ON THE DIVERGENCE OF PREDICTION AND UNDERSTANDING

In this section I employ the criterion of intelligibility presented above in order to ask whether some examples of neuroscientific models are intelligible. We will find that intelligibility is not an all or nothing property, that the models differ in their degrees of intelligibility, and such differences track the degree of understanding provided by the models. Even though Hempel was dismissive of the significance of understanding for philosophy of science, considered as a logic of the scientific method, he did assert that theories which satisfy the conditions for deductive-nomological explanation do also provide understanding:

> The argument shows that, given the particular circumstances and the laws in question, the occurrence of the phenomenon *was to be expected;* and it is in this sense that the explanation enables us to *understand why* the phenomenon occurred. (Hempel 1965:337, quoted in de Regt 2014:23-4)

In this remark, Hempel ties both explanation and understanding to successful prediction. The divergence of prediction and understanding, that I discuss below, should not be too surprising given that Hempel's account of explanation is no longer widely endorsed.

It is reasonable to assume that models which afford more accurate predictions of neural responses do so in virtue of being more accurate in their representation of the actual computation performed by the target neural system. The observation of a divergence of prediction and understanding suggests that these two epistemic benefits of science cannot always be served by the same means, namely that of representing the target system in the most accurate way possible. I will return to this matter in Section 4.

2.1 Decoding the Motor Cortex[13]

The trend I describe in this section is for decoders of the motor cortex to become less intelligible and more opaque as modelling technology has progressed. Here, the relevant function – the encoding model embedded in the decoder -- is a

[13] There are many more varieties of decoder than I can review in this brief section. The three classes I discuss here are prominent in the field and indicative of the trend I am investigating. For review of a wider range of decoders see Koyama et al. (2010) and Li (2014).

mapping from parameters of an intended movement (e.g. velocity of arm) to neural responses:

(2) *Response = f (intended movement)*

Given the unresolved question of what motor cortex neurons encode or represent -- if anything[14] -- it is a striking fact that a linear model relating neuronal firing to intended direction of movement, and a simple aggregative pooling rule, was used to decode M1 activity for nearly three decades. The *population vector algorithm* (PVA) (Georgopoulos, Schwartz, and Kettner 1986) makes the false assumptions, (1) that firing rate of typical neurons varies as a cosine function with intended direction of movement,[15] and (2) that the distribution of preferred directions is uniform in M1. However, deleterious effects from the false assumptions do not arise in all conditions. As Koyama et al. (2010) report, the bias introduced by (2) is compensated for when the decoder is used to generate movement commands in real time. It is fair to characterize this first generation algorithm as a highly intelligible, representationally inaccurate but surprisingly useful model of motor cortex.

Because of noise introduced in the recording process, and the inherent trial-to-trial variability of neuronal responses, methods for smoothing the data play an important part in the success of a decoder. A substantial advance was made in this regard with the introduction of the Kalman Filter (KF) in BMI research by Wu et al. (2006). KF decoders still posit a linear relationship between neural activity and output kinematics but they use Bayesian methods such that the predicted movement is informed by a prior expectation of the trajectory, itself continually updated as decoding proceeds. This smoothing counteracts the effect of noise in the data that would, if uncorrected, lead to jittery and misdirected motor output. This generation of models is currently employed in human trials for invasive BMI.

A third, entirely different approach to the decoding problem is to use machine learning -- training an artificial neural network to associate neural data with intended movements without building an explicit encoding model, or making assumptions about what M1 neurons represent. In work from the Shenoy group,

[14] For the view that M1 neurons do not represent anything, see Shenoy, Sahani, and Churchland (2013), and see Chirimuuta (2020) for further discussion.
[15] This is the "cosine tuning" encoding model embedded in the PVA decoder.

a *recurrent neural network* (RNN)[16] is shown to out-perform a KF decoder, with respect to the measure of minimising time taken for the user of the BCI to reach the targets. This new approach is motivated by an appeal to the greater realism that comes with a decoder sensitive to non-linear mappings between neural activity and intended movements:

> A standard decoder in BMI systems, the VKF [*velocity* KF], has seen wide application and performs better than its static counterpart, the linear decoder, presumably due to the Kalman filter's ability to capture aspects of the plant dynamics in the kinematic data. However, due to the linearity of the Kalman filter, the power of the VKF must be limited in contexts where the relationship between the inputs and outputs is nonlinear. While the nature of motor representation in the pre-motor dorsal cortex (PMd) and motor cortex (M1) remains an open question, it seems likely that the relationship between neural activity in these areas and arm kinematics is nonlinear. Thus, it is appropriate to explore nonlinear methods when decoding arm kinematics from PMd/M1 activity. [citations omitted; (Sussillo et al. 2012: 1-2)]

The irony is that the enhanced realism of moving to a non-linear decoder cannot be cashed out as a new, more accurate *and* equally intelligible model of motor cortex. Barak (2017: 2) contrasts RNN's that are designed according to hypotheses about the mechanisms or computations responsible for the neural population's behaviour, with those that are trained to reproduce a mapping from inputs to outputs and are hypothesis free. The RNN decoder is of the latter sort and therefore is, as Barak (2017:3) puts it, "somewhat of a black box."

In a paper from the Shenoy group which compares the performance of an RNN decoder with the currently best performing KF ("FIT-KF"), and demonstrates the advantage of the RNN with respect to speed and accuracy of movement, and robustness in the face of day to day variation in neuronal responses, the realism of the nonlinear decoder is emphasised along with its *potential* benefits[17] for technological applications outside of the laboratory, due to its ability to utilise information contained in large neural datasets. As Sussillo and co-authors state:

[16] An artificial neural network with feedback loops. Barak (2017) is a useful guide to the use of RNN's in neuroscience.

[17] I emphasise "potential" because these models have not yet shown their worth in human trials testing everyday applications such as control of an iPad. One obstacle is the high computational demand of running an ANN to decode in real time, while developers of BMI's for patients are aiming at devices that are compact and portable.

> Using this historical data would be difficult for most BMI decoders, as they are linear. Linear decoders are prone to underfitting heterogeneous training sets, such as those that might be sampled from months of data. To overcome this limitation, an essential aspect of our approach is to use a nonlinear and computationally 'powerful' decoder (that is, one capable of approximating any complex, nonlinear dynamical system), which should be capable of learning a diverse set of neural-to-kinematic mappings. (Sussillo et al. 2016:2; citations omitted)

The successful utilisation of large datasets over theory-driven approaches to modelling the motor cortex would be another instance of the "unreasonable effectiveness of data" (Halevy, Norvig, and Pereira 2009). One can draw an analogy with the big data approach to translation, where algorithms trained on masses of pre-translated texts have been shown superior to older approaches based on hypotheses about natural language structure. The downside, for neuroscience, is that the model builders themselves have limited information and insight regarding how the enhanced performance is achieved. Thus Gao and Ganguli (2015:151) argue that the architects and users of the most advanced artificial networks cannot be said to understand their own creations because such models do not meet the condition for intelligibility introduced above: modellers are not able to give qualitative descriptions of the model's behaviour under different conditions prior to running through the simulations. For example, the RNN model builder would not be able to describe, qualitatively, the relationship between velocity of intended movement and neural firing rate, whereas this is easy to express for the first generation model (see Figure 1). And if the models are not intelligible, they cannot be expected to provide understanding of the neural code in the systems they represent. The example of motor cortex decoders is not an isolated case. The same trend from linear, intelligible and inaccurate models to non-linear, opaque but predictively accurate ones can be found in research on the visual cortex.

---------FIGURE 1 NEAR HERE----------

Figure 1. Illustration of how qualitative predictions of movement associated with neural population activity can be obtained from first generation, "simple model" – the *population vector algorithm*. This model assumes that every neuron fires most strongly for one preferred direction of movement. The predicted movement is the sum of represented directions, weighted by intensity of firing rate. Here one sees that a high firing rate from the 'downward' preferring neuron, and a moderate firing from the 'left' preferring movement leads to an overall predicted movement down and slightly to the

left. In contrast the "ML model", an RNN, makes no explicit commitments about the relationship between firing rate, neuronal tuning, and the movement predicted. The model user cannot describe, in qualitative terms, the relationship between recorded responses and predicted movement. [from Glaser et al. 2019, Figure 2, permission needed]

2.2 Modelling the Visual System

In an interview, neuroscientist Adrienne Fairhall reflects on the unease prompted by this trend:

> A lot of work I and others have done in the past tries to extract coding models of data — for example, to try to fit a receptive field to predict an output. With these emerging methods to analyze high-dimensional data, rather than fit a receptive field, you train a randomly connected recurrent network to produce a certain kind of output. It's different than a simple receptive field model. You often get more accurate predictions of what the system will do. But maybe you're giving up an intuition about what's going on, so we end up building network solutions that we don't really understand.[18]

In visual neuroscience the encoding model is typically characterized as a receptive field (RF) describing the relationship between visual stimulus parameters and intensity of neural response, where: [19]

(3) *Response = f (stimulus)*

As in the motor cortex case, the first generation of models of retinal ganglion cells (RGC's) and primary visual cortex (V1) "simple cells" were highly intelligible and surprisingly effective: it was supposed that these neurons perform a linear sum of light falling in inhibitory and excitatory portions of their receptive fields, and that this sum is converted into a spike rate by an output non-linearity. Hence these models are sometimes referred to as "linear/nonlinear" (LN) models. Such models make fairly accurate predictions of responses to very simple stimuli displayed in the laboratory, such as dots or bars of light, but fail to predict responses to natural

[18] https://www.simonsfoundation.org/2018/01/02/the-state-of-computational-neuroscience/

[19] See Chirimuuta and Gold (2009) on the RF concept and a more detailed discussion of first and second generation work in visual neuroscience described here. See Carandini et al. (2005) for a useful review of the strengths and weaknesses of these models.

images or any complex artificial stimuli that elicit responses from a group of neurons with more than one kind of tuning preference.[20]

This limitation indicated the need to take inhibitory interactions between neurons with different tuning preferences into account. The second generation encompasses such interactions using relatively simple formulae to summarise the effects of inhibition between simple cells – the Normalization Model (Heeger 1992); or correlations between RGC responses -- the Generalized Linear Model (Pillow et al. 2008). However, these models have again been found wanting in their ability to predict responses to natural stimuli (David, Vinje, and Gallant 2004; Heitman et al. 2016).

The problem of accurately predicting responses to natural images, such as photographs and movies, has been solved by the third generation of models, *convolutional neural networks* (CNN's). This is the class of model responsible for the recent breakthroughs in computer object recognition. Unlike the recurrent neural networks discussed above, the architecture of these is entirely feedforward. A paper from Surya Ganguli's lab (McIntosh et al. 2016), describes a CNN trained to find the stimulus-response mapping for RGC data where the stimulus was 25 minutes' worth of moving images--either natural stimuli (recordings of real scenes) or white noise movies, like the "snow storm" on an untuned analogue television. It is important to emphasise that these models are able to predict responses to novel stimuli and have not merely fit the training data. Particularly impressive is that the CNN trained on data collected when neurons were stimulated with white noise makes fairly good predictions of neuronal responses to natural stimuli, whereas the previous generations of models did not generalise in this way.

Likewise, Cadena et al. (2019) trained CNN's to predict the responses of V1 neurons. Their networks outperformed the best of the second generation models. They observe that,

> [r]ecent advances in machine learning and computer vision using deep neural networks ('deep learning') have opened a new door to learn much more complex non-linear models of neural responses (Cadena et al. 2019:2).

However, this invites the question of who is "learning" these complex nonlinear models. Lacking *functional transparency* (Creel forthcoming), the mathematical relationship learned by the CNN, between visual stimuli and neuronal responses, is not made explicit to the human model builder. I will argue in the next section

[20] See Demb and Tolhurst sections in Carandini et al. (2005).

that this is one barrier to such models providing understanding of the visual cortex.

3. THE TRADE-OFF

The notion that two or more of the desiderata held important for theories or models in science trade off against each other and cannot be simultaneously optimised has occurred elsewhere in the philosophy of physics and biology (Levins 1966; Cartwright 1983; Cushing 1991). To make the case that a trade-off between prediction and understanding occurs in this branch of neuroscience, where models are built to represent computations occurring in neural systems, I will first say more about what the lack of intelligibility of the most predictively powerful models consists in, and how predictive power and intelligibility relate to one another. Then in Section 3.2 I address the question of how future developments in ANN modelling may or may not increase the intelligibility of the networks, arguing that even if they do become more transparent they will remain relatively less intelligible than the previous generations of models, such that the trade-off between prediction and understanding will still obtain. This supports my claim that the trade-off is not contingent on the current under-developed state of methods for reverse engineering ANN's.

3.1 Sources of Intelligibility

Above I observed that the most predictively powerful models failed the test of intelligibility: they did not allow scientists to make the kinds of qualitative predictions of neuronal responses afforded by the earlier models. I will now say more about what this difference in intelligibility consists in, discussing four characteristics of the earlier models that led to their greater intelligibility: (1) visualisability, (2) theoretical articulation, (3) linearity, and (4) functional transparency.

The characteristic of visualisability features prominently in discussions of the intelligibility of quantum mechanics (de Regt 2017, chapter 7). In physics the value of visualisability is still controversial. Fortunately, for our discussion, the significance of visualisability in neuroscientific modelling does not depend on the outcome of the debate in physics. By the characteristic of visualisability, I mean here specifically whether or not the model is accompanied by a picture of neural

coding, such as a "circuit" or "wiring diagram", which indicates how neurons could be connected together in order to generate the responses described quantitatively by the model. A well known example from visual cortex modelling is given in Figure 2. The coding schematic of the "simple model" of Figure 1 would also count as a visualisation of the population vector algorithm.

----------FIGURE 2 NEAR HERE----------

Figure 2. Diagram of the normalisation model of primary visual cortex (V1). Circles depict linear weighting functions, with excitatory (bright) and inhibitory (dark) subregions. Arrows represent excitatory connections, while connections terminating in curved lines are inhibitory. [Figure 1 from Heeger (1992), permission needed]

This kind of visualisability facilitates qualitative predictions of the model's behaviour, and it also gives researchers a hypothesis as to the physiological and anatomical basis of the neuronal responses, which lends itself to experimental investigation, opening out paths of investigation. ANN's used to generate decoding and encoding models are not visualisable in this sense -- they do not offer visualisable pictures of neural circuits implementing the functions that they compute. Remember that in these cases, the ANN does not represent anything in the anatomy of the neural system under investigation. Rather, the ANN is used purely as a mathematical instrument for learning the complex function relating stimulus conditions to neural responses. Therefore, in our cases the ANN architecture itself cannot be used as a visualisable picture of the target visual or motor system.

The next point of comparison is whether a model has articulated theoretical assumptions or is hypothesis free. This came up in the discussion of M1 decoders, where early generation models made assumptions about the neural code in that region, whereas machine learning based decoders do not make such assumptions. Such assumptions produce intelligibility because they give model builders information about the conditions under which different model outputs will occur, thus allowing for qualitative predictions. The early generation visual models also came with explicit assumptions about the code – e.g. that retinal ganglion and simple cells have the function of summing quantity of light falling within their excitatory fields. Again, this leads to qualitative predictions as to the stimulus conditions under which a neuron will give the strongest response. The important matter of the empirical testing and falsity of these assumptions will be brought up in Section 4.2.

The assumptions of the early M1 and V1 models – respectively, that the intended movement is governed by a sum of responses of individual neurons with different tuning preferences, and that simple cell response represents the sum of light falling into the excitatory field – relate to the linearity of these models. In other words, the assumption that the neural code in these regions basically involves a summation is what motivates their being modelled as performing linear computations. The intelligibility of (essentially) linear models is high because of the proportionality relationship between the input and output of the model. This makes for easy qualitative predictions of model behaviour. Qualitative prediction becomes harder for nonlinear models, but not impossible if the model is functionally transparent.

On the definition of Creel (forthcoming), a computer programme has *functional transparency* if it is possible to know the algorithm that the programme instantiates. Now, to say that the ANN models under discussion lack functional transparency I must reframe this definition slightly. While in these cases the algorithm that initially makes the model – the layered architecture, number of nodes, the learning rule – may be fully known to the modeller, what is not known is the mathematical function mapping inputs to outputs in the trained model. I am applying the notion of functional transparency to this function, for the trained model, which is taken to be an approximation of the computation performed by the actual neural system. In contrast to the ANN's, the earlier hand-coded models have full functional transparency. Since this function is the "scientific product" most significant to the modeller's effort to understand the neural code, the lack of transparency will be an obstacle to the use of ANN's for increasing scientific understanding. For networks of sufficient size and complexity to achieve the performance described in my cases, there is not currently a method for 'recovering' the function from the trained network, though illuminating analyses are possible for very small ANN's (Beer and Williams 2015). Is this a permanent obstacle to the intelligibility of such predictively powerful models? In Section 3.2 I say more about the contingency and context-sensitivity of these obstacles to understanding.

The conclusions we can draw now are important for explication of the trade-off. We have said that the greater predictive power of ANN's is based on their achievement of a much closer approximation to the complex non-linear function computed by the actual neural system than possible with the simple hand-coded models. The thing now to recognise is how this capacity is inherently related to

their deficiency with respect to the four characteristics of intelligibility. A model that uses machine learning to achieve a close approximation to a very complex, nonlinear neural computation will not be as functionally transparent as a hand coded model. It will obviously not be linear. And so long as the approximation to the neural computation remains implicit in the trained model it will not be possible to relate it to theoretical assumptions about the neural code, or to visualise the coding scheme in a wiring diagram. Conversely, the traditional modelling methods that score high on the four characteristics of intelligibility, only achieve this by remaining inaccurate in the approximation of the actual neural computation. Thus these models of neural systems are either very intelligible, or predictively accurate, but not both.

3.2 Opening the Black Box?

One objection to the claim that this the trade-off obtains is to argue that it is just a temporary problem because the use of ANN's in neuroscience of this scale and complexity is quite new. One might hope that with further research and practice using them, such models will become as intelligible as the traditional sort. As mentioned above, the criterion of intelligibility is a context-sensitive one, which leaves it as an open possibility that one and the same ANN could be intelligible given a different background context of mathematical methods, neuroscientific concepts, and scientists' experience with modelling methods. Ultimately, I will argue that even if there is some increase in intelligibility the ANN's will remain relatively less intelligible than their hand-coded counterparts, and so the trade-off will persist.

"Explainable AI" (XAI) is much discussed, but most of the focus is on the use of machine learning as a tool for prediction where decisions have immediate consequences for citizens and society (Zednik 2019). Computer scientist Cynthia Rudin has argued that in such contexts there are cases of equivalent accuracy being reached by other kinds of models, including linear models, that are much more interpretable than standard deep neural networks (e.g. Rudin and Radin 2019), and also that for image classification CNN's with "interpretability constraints" can achieve "comparable accuracy" to standard CNN's (Chen et al. 2019). While these are important findings, I emphasise that they are not very relevant to my cases of the use of ANN's as a tool in scientific discovery. They are certainly not counter examples to my claim that a trade-off between prediction and understanding occurs for neural decoding models. One obvious difference is

that the kind of interpretability useful in image classification – the model's output of what visual features influenced its classification – is not applicable to these cases in neuroscience.

It is more relevant to consider work on model interpretation within science. For RNN's used in neuroscience, methods have been developed to reverse-engineer trained networks in order to understand which of its features are responsible for its arriving at the solution to the task.[21] This yields some degree of functional transparency – some knowledge of the process transforming inputs to outputs in the trained network, but not full recovery of the function. Regarding theoretical articulation, Omri Barak (2017:1) argues that even though a trained RNN is hypothesis free by design, the process of reverse engineering them can lead to the generation of "complex, yet interpretable, hypotheses" about how real neural circuits perform their tasks.

Work on visualisation of information processing within CNN's is ongoing (Buckner 2018). It should also be emphasised that such visualisations cannot deliver functional transparency of the sort found in hand-coded models, as the procedure falls well short of the recovery of the function computed by the trained network. It is possible (though unlikely) that a procedure will be invented for making such functions explicit, so it is worth asking what the implications would be for the intelligibility of the model. I contend that even if we could write down by hand the equations embedded in the trained ANN's employed in my case studies, those models would still be far less intelligible than their low-tech predecessors, because they would be nonlinear and contain very many more terms than the ones occurring in the traditional models. Eyeballing an equation of such complexity would not give the neuroscientist the same qualitative sense of how adjustment of parameters or variables would make a difference to the behaviour of the system, and would not readily be associated with circuit diagrams of the brain area. For this reason I conclude that even for an ANN which has undergone an ideal degree of reverse engineering – whose internal layers have been visualised, and whose input-output function has been rendered explicit – it will still be relatively less intelligible than a hand-coded model and for this reason the trade-off between prediction and understanding for neural decoders will not go away even if reverse-engineering of ANN's progresses far beyond what is possible today.

[21] Sussillo and Barak (2013); for discussion see Chirimuuta (2018b: §4).

3.3 Scope of the Trade-Off

I have just argued that the presence of the trade-off will be a permanent feature of research in this area of neuroscience, not one that will just go away as progress is made in building more interpretable ANN's. Although ANN's may well become more intelligible to the scientists using them, they will always be less intelligible than their low-tech counterparts, if for no other reason than their much greater mathematical complexity. At the same time, there are cases elsewhere in science where the same model is both highly intelligible and predictively accurate. We know that the trade-off does not hold universally. This naturally raises a question about the scope of the trade-off as I have described it. Is it no more than a one-off instance in a small sub branch of computational neuroscience, or should we expect it to pop up elsewhere as the use of machine learning in science becomes common?

My expectation is that the trade-off will occur beyond these two case studies, in scenarios where the same basic problem structure occurs, which could be in the biological, physical or social sciences. Namely, scenarios in which the target system comprises a complex set of nonlinear dependencies that can only be poorly approximated by linear functions or the more simple nonlinear functions that can be arrived at through traditional modelling methods. Accurate representation of these complex relationships is a requirement for accurate prediction, but the complexity of the model able to achieve this more accurate representation will make it less intelligible, for the reasons given above. Moreover, the scenario will bear an important similarity to my neuroscience cases when scientists are striving to predict the system's behaviour in naturalistic conditions, where the challenge is to make predictions outside of the contrived experimental conditions that artificially simplify the system's behaviour. If the decoding challenge discussed above had not had the ambition to predict visual and motor system behaviour unfettered by the simplifying constraints of the laboratory (where restriction can be made to artificial visual stimuli and stereotyped motor responses), the early generation models may well have been regarded as perfectly adequate. This leads to the expectation that the trade-off is more likely to be felt in branches of research that put a premium on predictive success outside of controlled conditions, because the ultimate goal of research is to develop applications that work "in the wild". Biological research aiming at medical application is an obvious example here.

4. IMPLICATIONS FOR EXPLANATION AND NON-FACTIVE UNDERSTANDING

In this section I take a deeper look at the explanatory status of the ANN's presented in the two case studies, arguing in Section 4.1 that they do provide certain kinds of explanations but that this does not in itself allow them to confer understanding. In Section 4.2 I explain why the existence of the trade-off lends endorsement to non-factive accounts of scientific understanding.

4.1 Explanation With and Without Understanding

Up to this point I have been silent on the explanatory status of the various models discussed above. I now present a discussion of the kinds of explanations afforded by the different kinds of models, and how this lines up with intelligibility and ultimately to their ability to increase understanding. In contrast to de Regt's account, where it is assumed that any pairing of an intelligible physical theory and explanatory model will yield understanding of the system described, I point out that in neuroscience there are models which fulfil some common criteria for being explanatory, but are nonetheless not intelligible and therefore do not, by themselves, enhance the model-builder's understanding of the target system.

In our cases, the *broad* explanandum phenomenon is the response of neurons in the motor cortex, or visual system, either to a range of visual stimuli, or under specific conditions of motor intention. Classes of neurons, such as V1 simple cells, exhibit similar patterns of activation, so the explanandum phenomenon can also be thought of as the behaviour of the neuronal type rather than the responses of an individual neuron. Because all of the models target the mathematical relationship between external states and neural responses (Equation 1), not the biological mechanisms giving rise to the neurons' responses (i.e. physiological mechanisms of the target neuron), nor the causal process leading up to the responses (e.g. causal chain from stimulus, to eye, to visual cortex), the models offer computational explanations, of the sort discussed in Chirimuuta (2014; 2018a; 2018b), rather than the constitutive or aetiological explanations discussed in the literature on mechanisms in neuroscience (Craver and Darden 2013).

As argued in those previous publications, I submit that the first and second generation linear-nonlinear models of visual neurons provide *efficient coding explanations* of the responses: they specify a mathematical function potentially computed by the neurons, and offer information theoretic reasons why it would

be efficient to process visual information in this manner. They provide answers to the more specific questions (explananda), "*what is computed by the neurons, and why this function rather than another?*" Similarly, the linear encoding models at the heart of the first and second generation motor cortex BCI's answer the "what is computed?" question, and also suggest reasons why the brain would represent movements in this manner -- though for reasons not so much due to efficiency of processing but more in terms of the channelling of the information relevant to governing downstream neurons and muscles.

I have argued that such models often fail the necessary condition for constitutive mechanistic explanation -- "models-to-mechanism mapping" (3M) (Kaplan and Craver 2011) – while satisfying one condition for interventionist explanation, the ability to answer "what-if-things-had-been-different-" or "w-questions" (Woodward 2003). If we now turn to the ANN's used for decoding, we find that they all fail the 3M condition because there is no sense in which the nodes and connections of the ANN should be thought of mapping onto structures in the actual brain. As emphasised above, the purpose of building the network is not, here, to represent biological networks but to use the computational power of the ANN to find the function that maps external states to neural responses.[22] So neither the ANN's nor the traditional models should be thought of as offering aetiological or constitutive mechanistic explanations because they are not intended to represent the causal processes which generate the neurons' responses.

It can be said of the networks that go beyond the trained data and make accurate generalisations to new cases that they satisfy the condition for interventionist explanation. For example, they can answer w-questions by reporting what the responses would have been if other stimuli had been presented. Moreover we could say, for instance, that the network for predicting a simple cell's response is exhibiting a relationship of causal dependency between the arrangement of pixels in a visual stimulus and the neuron's firing rate. It is just that it is not revealing anything of the causal process that leads from the stimulus, via the early visual system, to the neuron's output; and that would not be a condition for interventionist explanation in any case. However, if one reads Woodward (2003) as proposing something more stringent – that explanatory theories and models

[22] Again, I emphasise that the cases presented above form the target of my analysis; I am not making the general claim that ANN's are *never* constrained by biological plausibility and so cannot offer explanations by this route (see Stinson forthcoming).

must e*xplicate* a dependency relationship – then we should say that ANN's fall short of interventionist explanations. I leave this as an open question.

In terms of Hempel's *inductive-statistical* category of explanation, today's AI's are stunningly successful. They are far better at making inductions on the basis of statistical regularities than the hand-crafted models of the earlier generations. This is likely because the networks are sensitive to subtle patterns in the neural data which appear only as noise to a human building a model from computational first principles and observation of datasets. Hempel (1966: 832) writes that explanation is achieved "by exhibiting the phenomena as manifestations of common, underlying structures and processes". This is quite a good description of the achievements of some AI in neuroscience. For example, the LFADS data smoothing algorithm employs an RNN, taking noisy neurophysiological data and learning the latent structure in the dataset which can then be used to generate a "cleaned up" version of the data (Pandarinath et al. 2017). It can rightly be said to show how the noisy recorded data are manifestations of the underlying structures of the neural population activity; the catch is that the patterns this model latches onto are not made available to the human user because they remain implicit in the trained network.

If one followed Hempel, one would be tempted to declare the problem of understanding neural coding in primary visual and motor cortex solved. The task was to find a function that very accurately maps external variables to neural responses (Equations 1-3). Such functions have been found, though they are implicit in the neural networks. All of the candidate functions offer answers to the w-question "what would the response be if the input were…..?", but only the AI solutions have met the neuroscientists' own standards for predictive accuracy. On the Hempelian view this ought to count as having an understanding of these brain areas, and what is missing is merely the subjective feeling of comprehension.

In response, I argue that these Hempelian explanations are insufficient for understanding because they tell you what is to be expected, but not why.[23] Of particular importance is the ANNs' failure to answer the *contrastive* question of the form, "why this encoding function and not another?", or to provide the

[23] Following Khalifa (2017:2) I take it that "explanatory understanding" is equivalent to "understanding why", such that understanding of a system enables one to explain why certain things happen.

requisite information on the basis of which this question might be answered.[24] So long as the functions which solve the prediction problem remain embedded in the networks, they cannot be analysed in relation to information theoretic principles, or hypotheses about the neural code. We do not know *what*, according to the network, the visual or motor cortex neurons are computing and this means that we are left in the dark about the significance of the ANN's discovery for our broader theories of neural function.

Table 1: Comparison of Models

	FIRST & SECOND GENERATION	*ARTIFICIAL NEURAL NETWORK*
PREDICTION	Fails outside simple cases	**Impressive across cases**
EXPLANATION	Not mechanistic; Maybe I-S; **Interventionist; Efficient Coding**	Not mechanistic; **I-S;** Maybe interventionist; Not efficient Coding
INTELLIGIBILITY	**Display four characteristics of intelligibility. Enable qualitative predictions.**	Lacking in four characteristics of intelligibility. Do not enable qualitative predictions.
UNDERSTANDING	**Intelligible; Provide explanations that answer "WHY?" questions.**	Not intelligible; No explanations that answer "WHY?" questions.

Table 1 summarises the comparison of AI and traditional models in terms of their ability to predict, explain, and offer understanding.

4.2 Is Understanding Factive?

To recap, a longstanding goal of computational neuroscience has been to produce predictively accurate models of neural responses to a wide range of

[24] Elsewhere I present the case that "efficient coding explanations" offer answers to this kind of question (Chirimuuta 2014; 2018a; 2018b). Similarly, Fairhall (2014:ix) writes, "[receptive field] theory has addressed not just what is encoded, but why the encoded features may assume the form they do. Two key principles have emerged: that these features may provide an efficient way to represent the specific statistical structure of the natural world, and that neural representations are sparse, in the sense that any natural input can be represented by the activation of relatively few neurons."

external conditions, which also confer understanding of the systems modelled. I have argued that this ambition has not been realised by any single kind of model: models simple enough to be intelligible give false descriptions of the function computed by the neurons such that their predictions fail beyond a very limited range of conditions; models sophisticated enough to closely approximate the complex nonlinear functions computed by actual neurons, and hence give very accurate predictions across a range of conditions, are not intelligible to the scientists. I have argued that even if the AI models become more intelligible with time, they will always be relatively less intelligible than the hand-coded ones, which means that the trade-off will still obtain.

An objection to my account poses the question: *how is it that a less realistic model can be said to provide more understanding?* The intelligible models are false of the neural systems so, one might object, it is a mistake to say that they confer understanding. This line of objection presupposes a *factive* notion of understanding, one that treats understanding as resting on the scientist having the relevant true beliefs (Khalifa 2017:155). This indicates that a *non-factive* account of understanding is required to make sense of my case studies and the finding of the trade-off.

The non-factivist *denies* that understanding requires belief in true or approximately true explanations of the phenomenon (Khalifa 2017:156). The core intuition of non-factivism is that theories and models that confer understanding are a compromise between the mind-boggling complexity of nature and the limited human capacity to make sense of complex patterns of phenomena. A model of the brain that just presented *The Truth* of the brain (in the sense of a representation that copies some or all of its features) would be no more comprehensible to us than the brain itself. Therefore substantial abstraction (simplification) and idealisation (distortion) are the departures from the truth that are the necessary ingredients of intelligible models. As de Regt summarises:

> an approximately true description of the system is no precondition for understanding; on the contrary, if one wants to understand a complex system it is often advisable to abandon the goal of a realistic description. Typically, representations that are closer to the truth are less intelligible and accordingly less useful for achieving scientific understanding. (de Regt 2015:3789; cf. Elgin 2004; Potochnik 2017:chap. 4).

While there are important objections to non-factivism (Khalifa 2017: chapter 6),[25] the findings of this paper suggest that the relationship between a model's representational accuracy, and its capacity to render nature in a way comprehensible to us, is contingent at best.

It was mentioned above that the assumptions made by the early generations of models were found to be incorrect on empirical testing. For example, it turns out that all directions of movement are not represented by an equal number of neurons in M1, as assumed by the PVA model (Koyama et al. 2010). However, employment of these models carried on even after these empirical "falsifications". Again, this is consistent with the non-factivist point that the aims of science are not always to be pursued by making theories and models more realistic. Or as Elgin (2017) puts it, models with false assumptions are often "true enough" for certain purposes – such as decoding under some conditions, and generating an understanding of the target system.

5. CONCLUDING REMARKS

In this paper I have argued that research on decoding in computational neuroscience reveals a trade-off between predictive accuracy and the ability of the models to confer understanding. If I am right, progress in accuracy of prediction will not be accompanied by progress in understanding the brain. Faced with the criticism that the use of machine learning to model the brain does not yield an acceptable level of understanding, some neuroscientists have suggested that understanding be redefined as ability to predict and control, and that understanding can be operationalised as the ability of the experimenter to control the brain (Bashivan, Kar, and DiCarlo 2019: 1). By articulating a fuller notion of scientific understanding, contrasting it with prediction, and applying it to cases in computational neuroscience, I hope to have shown why it would be unsatisfactory to opt for semantic revision.

I have concluded that a non-factivist account of understanding – one which takes understanding to be provided by models that reduce natural complexity down to a humanly manageable size via abstraction and idealisation – is reinforced by the finding of the trade-off. It should not be surprising that intelligibility, a human-relative virtue of models, is compromised when models of natural systems are

[25] Responses are in preparation for a future publication.

learned by algorithms instead of being devised by humans. An important question for the scientific community is whether a place must be retained for models which retain intelligibility, even if their instrumental utility falls short in comparison to high-tech rivals. In other words, the question of the value of understanding as an end in itself, not as means towards prediction and control, is presented to us by the trade-off outlined in this paper.

ACKNOWLEDGEMENTS

I am much indebted to audiences at the Philosophy departments at the University of Edinburgh and University of Birmingham, Center for Philosophy of Science at the University of Pittsburgh, and the 2018 meeting of the APA (Pacific) for many provocative comments and suggestions. I am also much appreciative of lively discussions the VISCOG group at Carnegie Mellon University. Finally, I thank Colin Allen and Jim Woodward for comments on an early version of this manuscript, and two anonymous reviewers for many helpful criticisms.

References

Barak, O. 2017. 'Recurrent neural networks as versatile tools of neuroscience research', *Curr Opin Neurobiol*, 46: 1-6.

Bashivan, Pouya, Kohitij Kar, and James J. DiCarlo. 2019. 'Neural population control via deep image synthesis', *Science*, 364: 1-11.

Beer, R.D., and Paul L. Williams. 2015. 'Information Processing and Dynamics in Minimally Cognitive Agents', *Cognitive Science*, 39: 1–38.

Buckner, Cameron. 2018. 'Empiricism without magic: transformational abstraction in deep convolutional neural networks', *Synthese*, 195: 5339 - 72.

———. 2019. 'Deep Learning: Philosophical Issues', *Philosophy Compass*, 14: e12625.

Cadena, Santiago A., George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker. 2019. "Deep convolutional models improve predictions of macaque V1 responses to natural images." In *PLoS Comput. Biol.*, e1006897.

Carandini, M., J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust. 2005. 'Do we know what the early visual system does?', *J Neurosci*, 25: 10577-97.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie* (Oxford University Press: Oxford).

Chen, Chaofan, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. 2019. ' This Looks Like That: Deep Learning for Interpretable Image Recognition', *Advances in Neural Information Processing Systems*, 32.

Chirimuuta, M. 2013. 'Extending, changing, and explaining the brain', *Biology & Philosophy*, 28: 613-38.

———. 2014. 'Minimal Models and Canonical Neural Computations: The Distinctness of Computational Explanation in Neuroscience', *Synthese*, 191: 127-53.

———. 2018a. 'The Development and Application of Efficient Coding Explanation in Neuroscience.' in J. Saatsi and A. Reutlinger (eds.), *Explanation Beyond Causation* (Oxford University Press: Oxford).

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

———. 2020. '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).

———. forthcoming. 'Your brain is like a computer: function, analogy, simplification.' in Fabrizio Calzavarini and Marco Viola (eds.), *Neural Mechanisms: New Challenges in the Philosophy of Neuroscience* (Springer: Berlin).

Chirimuuta, M., and I. Gold. 2009. 'The embedded neuron, the enactive field?' in John Bickle (ed.), *Handbook of Philosophy and Neuroscience* (Oxford University Press: Oxford).

Churchland, Patricia Smith, and Terrence J. Sejnowski. 2016. 'Blending computational and experimental neuroscience', *Nature Reviews Neuroscience*, 17: 667-68.

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

Craver, Carl. 2010. 'Prosthetic Models', *Philosophy of Science*, 77: 840-51.

Creel, Kathleen. forthcoming. 'Transparency in Complex Computational Systems', *Philosophy of Science*.

Cushing, James T. 1991. 'Quantum Theory and Explanatory Discourse: Endgame for Understanding?', *Philosophy of Science*, 58: 337-58.

Datteri, E. 2017. 'The Epistemic Value of Brain-Machine Systems for the Study of the Brain', *Minds and Machines*, 27: 287-313.

Datteri, Edoardo. 2008. 'Simulation experiments in bionics: a regulative methodological perspective', *Biology & Philosophy*, 24: 301-24.

David, S. V., W. E. Vinje, and J. L. Gallant. 2004. 'Natural stimulus statistics alter the receptive field structure of V1 neurons', *Journal of Neuroscience*, 24: 6991–7006.

De Regt, H. W. 2017. *Understanding Scientific Understanding* (Oxford University Press: Oxford).

De Regt, H. W., and D. Dieks. 2005. 'A Contextual Approach to Scientific Understanding', *Synthese*, 144: 137-70.

de Regt, Henk. 2009. 'Understanding and Scientific Explanation.' in Henk de Regt, Sabine Leonelli and Kai Eigner (eds.), *Scientific Understanding : Philosophical Perspectives* (University of Pittsburgh Press: Pittsburgh, PA).

———. 2015. 'Scientific understanding: truth or dare?', *Synthese*, 192: 3781–97.

du Bois-Reymond, E. 1874. 'The Limits of our Knowledge of Nature, Translated by J. Fitzgerald', *Popular Science Monthly*, 5: 17-32.

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

Elgin, Catherine Z. . 2017. *True Enough* (MIT Press: Cambridge MA).

Fairhall, A., and C. Machens. 2017. 'Editorial overview: Computational neuroscience', *Curr Opin Neurobiol*, 46: A1-A5.

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.

Gao, P., and S. Ganguli. 2015. 'On simplicity and complexity in the brave new world of large-scale neuroscience', *Curr Opin Neurobiol*, 32: 148-55.

Georgopoulos, A., A. Schwartz, and R. Kettner. 1986. 'Neuronal population coding of movement direction', *Science*, 233: 1416-19.

Glaser, Joshua I., Ari S. Benjamin, Roozbeh Farhoodi, and Konrad P. Kording. 2019. 'The roles of supervised machine learning in systems neuroscience', *Progress in Neurobiology*, 175: 126–37.

Halevy, A., P. Norvig, and F. Pereira. 2009. 'The unreasonable effectiveness of data', *IEEE Intelligent Systems*, 24: 8–12.

Heeger, D. J. 1992. 'Normalization of cell responses in the cat striate cortex', *Visual Neuroscience*, 9: 181-97.

Heitman, Alexander, Nora Brackbill, Martin Greschner, Alexander Sher, Alan M. Litke, and E. J. Chichilnisky. 2016.

Hempel, C. G. 1966. *Philosophy of Natural Science* (Prentice-Hall: Englewood Cliffs, NJ).

Hempel, Carl G. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science* (Free Press: New York).

Hooker, Giles, and Cliff Hooker. 2018. 'Machine Learning and the Future of Realism', *Spontaneous Generations*, 9: 174-82.

Kaplan, David Michael, and Carl Craver. 2011. 'The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective', *Philosophy of Science*, 78: 601-27.

Khalifa, K. 2017. *Understanding, Explanation, and Scientific Knowledge* (Cambridge University Press: Cambridge).

Koyama, S., S. M. Chase, A. S. Whitford, M. Velliste, A. B. Schwartz, and R. E. Kass. 2010. 'Comparison of brain-computer interface decoding algorithms in open-loop and closed-loop control', *J Comput Neurosci*, 29: 73-87.

Kriegeskorte, Nikolaus. 2015. 'Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing', *Annual Review of Vision Science*, 1: 417–46.

Levins, R. 1966. 'The strategy of model building in population biology.' in E. Sober (ed.), *Conceptual issues in evolutionary biology* (MIT Press: Cambridge MA).

Li, Z. 2014. 'Decoding methods for neural prostheses: where have we reached?', *Front Syst Neurosci*, 8: 129.

Lipton, Z.C. 2016. " The Mythos of Model Interpretability." In *arXiv:1606.03490*.

McIntosh, L.T., N. Maheswaranathan, A. Nayebi, S. Ganguli, and S.A. Baccus. 2016. 'Deep learning models of the retinal response to natural scenes', *Advances in Neural Information Processing Systems*, 29: 1369–77.

Naselaris, T., and K. N. Kay. 2015. 'Resolving Ambiguities of MVPA Using Explicit Models of Representation', *Trends Cogn Sci*, 19: 551-4.

Nicolas-Alonso, L. F., and J. Gomez-Gil. 2012. 'Brain computer interfaces, a review', *Sensors (Basel)*, 12: 1211-79.

Nishimoto, S., A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant. 2011. 'Reconstructing visual experiences from brain activity evoked by natural movies', *Curr Biol*, 21: 1641-6.

Omrani, M., M. T. Kaufman, N. G. Hatsopoulos, and P. D. Cheney. 2017. 'Perspectives on classical controversies about the motor cortex', *J Neurophysiol*, 118: 1828-48.

Pandarinath, Chethan, Daniel J. O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D. Stavisky, Jonathan C. Kao, Eric M. Trautmann, Matthew T. Kaufman, Stephen I. Ryu, Leigh R. Hochberg, Jaimie M. Henderson, Krishna V. Shenoy, Larry F. Abbott, and David Sussillo. 2017. "Inferring single-trial neural population dynamics using sequential auto-encoders." In *bioRxiv*.

Paninski, L., and J.P. Cunningham. 2018. 'Neural data science: accelerating the experiment-analysis-theory cycle in large-scale neuroscience', *Current Opinion in Neurobiology*, 50: 232–41.

Pillow, J. W., J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky, and E. P. Simoncelli. 2008. 'Spatio-temporal correlations and visual signalling in a complete neuronal population', *Nature*, 454: 995-9.

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

Rieke, F., D. Warland, R. R. Van Steveninck, and W. Bialek. 1999. *Spikes: Exploring the neural code* (MIT Press: Cambridge, MA).

Rudin, Cynthia, and Joanna Radin. 2019. 'Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition', *Harvard Data Science Review*, 1.

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

Stevenson, I. H., and K. P. Kording. 2011. 'How advances in neural recording affect data analysis', *Nat Neurosci*, 14: 139-42.

Stinson, Catherine. forthcoming. 'From Implausible Artificial Neurons to Idealized Cognitive Models: Rebooting Philosophy of Artificial Intelligence', *Philosophy of Science*.

Sussillo, D., and O. Barak. 2013. 'Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks', *Neural Comput*, 25: 626-49.

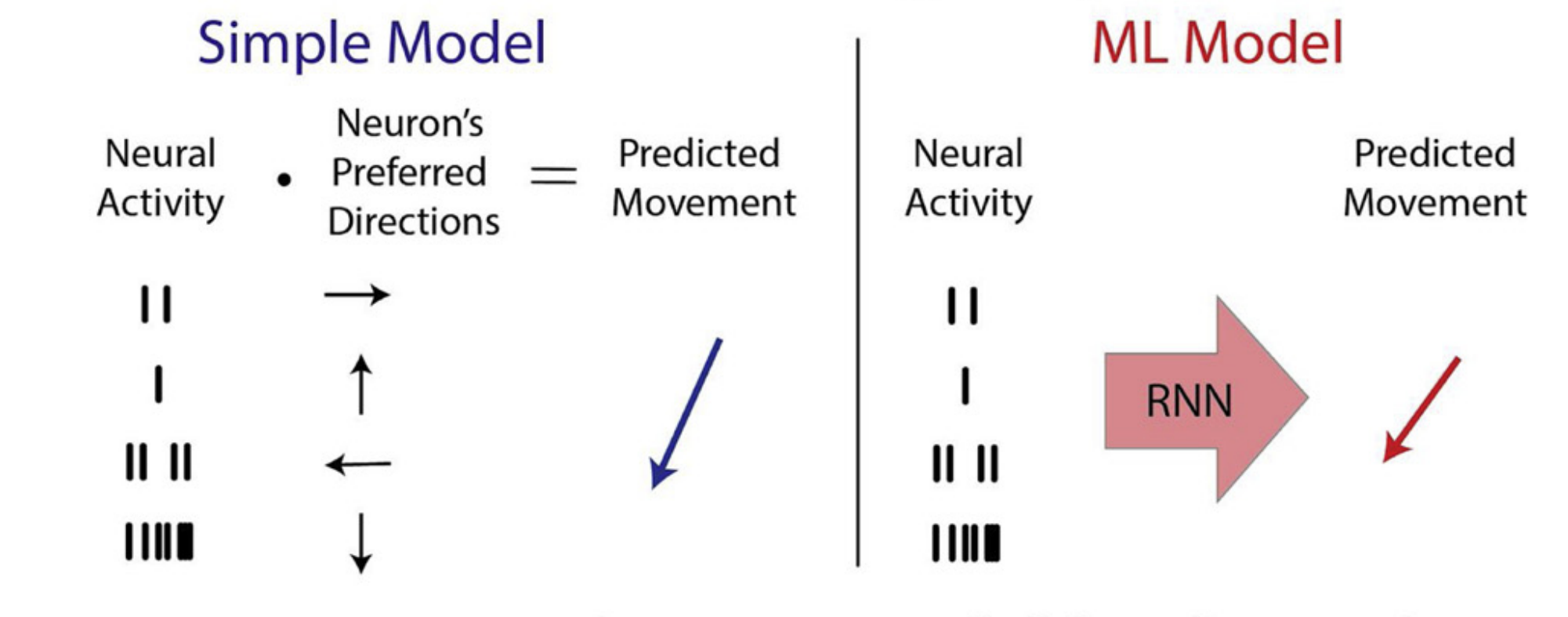
Sussillo, D., P. Nuyujukian, J. M. Fan, J. C. Kao, S. D. Stavisky, S. Ryu, and K. Shenoy. 2012. 'A recurrent neural network for closed-loop intracortical brain-machine interface decoders', *J Neural Eng*, 9: 026027.

Woodward, J.F. 2003. *Making Things Happen* (Oxford University Press: Oxford).

Wu, W., Y. Gao, E. Bienenstock, J. P. Donoghue, and M. J. Black. 2006. 'Bayesian population decoding of motor cortical activity using a Kalman filter', *Neural Computation*, 18: 80–118.

Yamins, D. L., and J. J. DiCarlo. 2016. 'Using goal-driven deep learning models to understand sensory cortex', *Nat Neurosci*, 19: 356-65.

Zednik, Carlos. 2019. 'Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence', *Philosophy and Technology*.


Simple Model
Neural Activity
Neuron's Preferred Directions
Predicted Movement
ML Model
Neural Activity
RNN
Predicted Movement

Figure