Prediction versus Understanding in Computationally Enhanced Neuroscience

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

ABSTRACT
The use of artificial intelligence instead of traditional models in neuroscience raises significant questions about the epistemic benefits of the newer methods. Here I argue that the benefit of providing the scientist with understanding of the neural system 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. In order to address objections to non-factivism, I recommend a modified account of the relationship between models and their targets in nature.

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. Patricia Churchland and Terry Sejnowski (2016:667) write that, possibly, "we are on the brink of a new era of collaboration between systems neuroscience and AI." They are optimistic that, "breakthroughs in machine learning … can be harnessed to find deep patterns and organizational features within" very large data sets [see also Paninski and Cunningham (2018)]. In contrast, Yves Fre⍰gnac (2017: 471) worries that,

> wishful thinking has replaced the conceptual drive behind experiments, as if using the fanciest tools and exploiting the power of numbers could bring about some epiphany.

This paper attempts to deliver a non-partisan analysis of the advantages and limitations of the use of machine learning in neuroscience, looking in particular at artificial intelligence based on connectionist networks (*artificial neural networks* or ANN's) 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 predictive power and the ability to enhance the scientists' understanding of the brain.

Philosophers such as Carl Hempel took understanding to be subjective and peripheral to philosophy of science,[1] 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,[2] 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 comprehensibility of AI and other complex computational models 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, presented in Section 2, between prediction and understanding afforded by computational models in neuroscience, I will argue in Section 3 that a non-factive account of understanding best suits the case in hand. Roughly speaking, non-factivists about understanding do not equate

[1] 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 being "naked prediction."

[2] This is a fascinating episode of intellectual history. See Finkelstein (2013: chapter 12) and references therein.

[3] Four books just 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.

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). In the final part of the paper I will respond to some important objections to non-factive accounts of explanation raised by Khalifa (2017).

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

[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.

> 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 has 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 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 also 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:

> 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)[5]

[5] Gao and Ganguli (2015:148) apply this criterion here:

"[T]he vague question of 'how the brain works' can be meaningfully reduced to the more precise, and proximally answerable question of how do the connectivity and dynamics of distributed neural circuits give rise to specific behaviors and computations? But what would a satisfactory answer to this question look like? A detailed, predictive circuit model down to the level of ion-channels and synaptic vesicles within individual

The idea is that to achieve comprehension, going beyond the bare ability to make accurate predictions, one must have a reliable sense of under what circumstances the classes of predicted phenomena obtain, 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 theory that affords visualisation of the natural system is typically more intelligible than a purely abstract theory because mathematical or concrete models can be constructed on the basis of such intuitive pictures (de Regt 2009:33; de Regt and Dieks 2005:155).

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 below, when I will consider if, with future developments, ANN's are likely to become more intelligible. I should also note that the criterion presented by de Regt has its origin in physics where there is a clear separation between fundamental theory and models. Within quantitative neuroscience, the terms "theory" and "model" are often used interchangeably; I will not be enforcing an artificial distinction here. In Section 2 I will use the criterion to ask whether my examples of neuroscientific theories/models are intelligible, and I will also consider what sort of explanations they afford.

1.2 Decoding the Brain

My focus is 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

neurons, while remarkable, may not yield conceptual understanding in any meaningful human sense. For example, if simulating this detailed circuit were the only way we could predict behavior, then we would be loath to say that we understand how behavior emerges from the brain."

represented by that activity. For example, Stevenson and Kording (2011: 140) state:

> 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 usually featured in the design of computer programmes at the heart of such devices. I refer to these programmes as *decoders*. In the case of motor-BMI's, these 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)*

One neuroscientist Garrett Stanley claims that successful decoding is an indicator of understanding brain activity:

[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)

> One clear litmus test as to whether we truly understand the neural code is whether we can tap into the activity of the neurons and make clear predictions about what is going on in the outside world or what is about to go on through the actions of the organism. (Stanley 2013: 259)

But as schematised in Equation 1, the encoding model is simply 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 really does help to advance any further epistemic goals, such as explanation and understanding. Philosophers of science -- in our post-positivist times -- may be surprised to find neuroscientists so accepting of the idea that there is an intrinsic link between prediction, explanation and understanding. Yet in the case of neural encoding models it has been reasonable to make this assumption. 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; 2017). Furthermore, 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).

Surprisingly, given the complexity of the brain, many of the models that have been predictively accurate (albeit in a limited range of experimental conditions) are simple linear functions (or linear models with straightforward nonlinear additions) that present no interpretative difficulties.[10] Until recently, in visual and motor neuroscience, there has been a fruitful co-alignment of the following goals: explanation, prediction, and understanding. The story I will tell in the next section is one of an emerging misalignment of these 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, arguing that there is a trade-off between the predictive accuracy and intelligibility of the models. In Section 3 I will elaborate on and defend a non-factive account of scientific understanding that helps to make sense of this trade-off.

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

2. THE DIVERGENCE OF PREDICTION, EXPLANATION AND UNDERSTANDING

De Regt's account of intelligibility was introduced in the previous section. In physics, intelligible theories are said to allow scientists to build models that explain natural systems. In the process of theory use and model building, the scientist comes to understand the target system. In neuroscience, a discipline which lacks this theory-model division of labour, one can speak of the quantitative representation of the neural system -- the 'theory/model' -- as providing explanations and thereby understanding. In this section I will present cases which show that theory/models which differ in their degrees of intelligibility provide different kinds of explanations of neural activity patterns, and such differences track the degree of understanding provided by the theory/models.[11]

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, explanation and understanding, that I discuss below, should not be surprising given that Hempel's account of explanation is no longer widely endorsed. However, it is important to note that models which afford more accurate predictions do so in virtue of being more accurate in their representation of the relevant relationships within the datasets derived from natural phenomena. Therefore the divergence between prediction, explanation and understanding suggests that all of these epistemic benefits of science cannot all be served by the same process of representing natural phenomena in the most accurate way possible. I will return to this matter in Section 3.

[11] 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.

2.1 Decoding the Motor Cortex[12]

The trend I describe in this section is for decoders of the motor cortex to become less intelligible and more opaque as neurotechnology has progressed. Here, the relevant function – the encoding model embedded in the decoder -- is a 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[13] -- 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 now believed false assumptions, (1) that firing rate of typical neurons varies as a cosine function with intended direction of movement,[14] and (2) that the distribution of preferred directions is uniform in M1. The deleterious effect of the false assumptions, especially (2), is clear in off-line decoding of neural data – where the algorithm is used to reconstruct the movement performed in an experiment conducted previously; however, in on-line (i.e. real-time) decoding the brain compensates for the bias introduced by (2) (Koyama et al. 2010). 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

[12] 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).

[13] For the view that M1 neurons do not represent anything, see Shenoy, Sahani, and Churchland (2013), discussed by AUTHOR (in press).

[14] This is the "cosine tuning" encoding model embedded in the PVA decoder.

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.[15] Krishna Shenoy's group made further improvements to the KF decoder by adding another calibration step along with "intention estimation" – adjusting the velocity prediction according to an estimate of the intended target.[16]

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 any explicit encoding model, or making any assumptions about what M1 neurons represent. In work also from the Shenoy group, a *recurrent neural network* (RNN)[17] is shown to out-perform a traditional KF, with respect to the measure of minimising time taken for the user of the BCI to reach the targets. This alternative approach is motivated by an appeal to the greater neural 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)]

[15] Note that all decoders employ some kind of smoothing method. A less sophisticated method is simply to average spike counts across a short window. Koyama et al. (2010) report that choice of smoothing method makes the biggest difference to the *on-line* performance of the decoders they tested.

[16] This decoder is known as the Re-FIT KF – "recalibrated feedback intention-trained Kalman filter" (Gilja et al. 2012). Fan et al. (2014) show that its performance gains are largely due to the intention estimation process. By Sussillo et al. (2016), the state of the art decoder in this class is taken to be the FIT-KF – i.e. intention estimation without recalibration.

[17] An artificial neural network with feedback loops. Barak (2017) is a useful guide to RNN's for neuroscientists.

The irony is that the enhanced realism of moving to a non-linear decoder cannot be cashed out as a new, accurate and equally intelligible theory 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 recent 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 not emphasised so much as its potential benefits for technological applications outside of the laboratory due, in part, to its ability to utilise information contained in large neural datasets.[18] As Sussillo and co-authors state:

> 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)

Amongst the practical advantages of the RNN decoder, these authors mention that it has the potential to decode successfully when recording conditions are less than ideal, and that the decoder could achieve stable performance without the need for long breaks for recalibration.

The downside is that the model builders themselves have limited information and insight regarding how this level of 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

[18] Another example of the "unreasonable effectiveness of data" (Halevy, Norvig, and Pereira 2009). One might draw an analogy with the big data approach to translation. Algorithms trained on masses of pre-translated texts have been shown far superior to older AI approaches based on hypotheses about natural language structure.

different conditions prior to running through the simulations. And if the models are not intelligible, they cannot be expected to provide understanding of the neural systems whose computations they represent.[19] 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.

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.[20]

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:[21]

(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

[19] I emphasise here that the ANN's discussed in my study are intended to represent neural computations, and *not* neural anatomy or physiology. E.g. the nodes and connections in an ANN for the motor cortex are not intended to represent a population of biological neurons and the connections amongst them. As discussed below, such models should not be thought of as representations of mechanisms.

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

[21] 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.

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 such as dots or bars of light, but fail to predict responses to natural images or any complex stimuli that elicit responses from neurons across the population that have a variety of tuning preferences.[22]

This limitation indicated the need to take interactions between neurons 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,[23] *convolutional neural networks* (CNN's).[24] Unlike the recurrent neural networks discussed above, the architecture of these is entirely feedforward. A paper from Surya Ganguli's lab (McIntosh et al. 2017), describes a CNN trained on RGC response data collected during the viewing of 25 minutes worth of movies (either natural images or white noise). It is important to emphasise that these models are able to predict responses to new stimuli and have not merely fit the training data. Particularly impressive is that the CNN trained on white noise makes fairly good predictions of responses to natural stimuli, whereas the previous generations of models did not generalise in this way.

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

[23] One caveat: the work discussed here is so new that it has not yet appeared in the peer-reviewed neuroscience journals. We can say that the problem has been solved, pending peer-review!

[24] These are more widely used than RNN's, and have been responsible for the recent breakthroughs in computer object recognition. The success of artificial visual systems has inspired visual neuroscientists to explore CNN's. E.g. Yamins and DiCarlo (2016), VanRullen (2017). This kind of AI is often referred to as "deep learning". For an accessible overview, see LeCun, Bengio, and Hinton (2015).

Likewise, Cadena et al. (2017) 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. 2017:2).

However, this invites the question of who is "learning" these complex nonlinear models. One way to describe the trend indicated by my cases is that the mappings schematised in equations (1-3) have evolved from functions that can be written down and pondered by a scientist considering the nature of neural computation, to functions that are embedded in a trained neural network, cannot be written down,[25] and so do not lend themselves to further scientific analysis.

To make a general point about the lack of intelligibility of the ANN's used in these neuroscience applications, it is worth noting that they do not offer a visualisable picture of any neural coding scheme. As mentioned in Section 1.1, theories which are readily visualisable, as an arrangement of interacting items, are typically more intelligible than non-visualisable theories. Now, while an ANN does consist of an arrangement of nodes and connections, these do not represent anything in the anatomy of the neural system that is the actual target of representation; rather, the ANN architecture is purely a mathematical instrument for learning the complex function relating neural external conditions to neural responses. Therefore, the ANN architecture cannot be used, in our cases, as a visualisable picture of the target visual or motor system. In contrast, the traditional models, especially of the visual system, are associated with very simple "wiring diagrams" regarding the inputs to the neurons being modelled. They come with a visualisable picture of neural coding which is itself suggestive of other models and coding schemes and lends itself to experimental investigation.

The lack of visualisability, and absence of an explicit function derivable from the trained network, are the two main reasons why the ANN's discussed in my cases

[25] My point here is that for ANN's of sufficient size and complexity to achieve the performance described in my cases, there is not currently any mathematical method for 'recovering' the function from the trained network. Illuminating analyses are possible for very small ANN's (Beer and Williams 2015). Of course, as Pat Churchland has emphasised (personal communication), this does not exclude the possibility that at some point in the near or distant future a mathematical breakthrough will occur which makes this procedure possible. I will say more about the contingency and context-sensitivity of these obstacles to understanding in Section 2.4.

are not intelligible models of neural systems. In Section 2.4 I will discuss possible ways that these obstacles to intelligibility might be overcome. For now, the conclusion to be drawn from my examples is that *in these cases* there is a trade-off between prediction and understanding. Neuroscientists have been able to build models of visual and motor cortex neurons' responses that are either intelligible, or very predictively accurate, but not both. The more one gains predictive accuracy across a greater range of stimulus conditions, the more one gives up intelligibility. I am not claiming that this trade-off occurs everywhere in science. Of course there are plenty of examples of theories and models that make very precise and accurate predictions while also being highly intelligible. This naturally raises the question about the scope of the trade-off as I have described it. However, it would apply beyond these two case studies to any areas of research in which ANN's, or other complex computational models, provide accurate empirical predictions but fall below the above-specified criterion of intelligibility. This could be in the biological, physical or social sciences.

2.3 Explanation With and Without Understanding

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 theory/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 explanandum phenomenon is the motor cortex or visual neuron's response 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, more broadly, 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 by Chirimuuta (2014 ; 2017;

2018), rather than the constitutive or aetiological explanations discussed in the literature on mechanisms in neuroscience (Craver and Darden 2013).

Following Chirimuuta, 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 questions, "*what is computed by the neurons, and why?"* 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.

Chirimuuta has 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, we find that they all fail the 3M condition because they are not constrained by biological plausibility, even of a very abstract kind. There is no sense in which the nodes and connections of the ANN should be thought of mapping onto structures in the actual brain. The purpose of building an artificial neural network is not, here, that of representing biological networks but of using the computational power of the ANN to find the function that maps external states to neural responses. Likewise, neither the ANN's nor the traditional models should be thought of as offering aetiological 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 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 describing 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 condition for interventionist explanation in any case. However, if one reads Woodward (2003) as proposing something more stringent – that explanatory theories and models must explicate 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, todays' 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 it 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. It could be argued that this ought to count as having an understanding of these brain areas, [26] 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.[27] In

[26] I thank Mark Sprevak for raising this objection.

[27] 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.

particular, the AI models fail to answer the question of the form, "why this encoding function and not another?", or to provide the requisite information on the basis of which this question might be answered.[28] 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 AI'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	**Explicit function; Qualitative picture of wiring/coding**	No explicit function; No qualitative picture of wiring/coding
UNDERSTANDING	**Intelligible; Provides 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. The trade-off between prediction and understanding comes about because of the different degrees of intelligibility of the models, and the different kinds of explanations they offer. It is not the first time that a trade-off between the virtues of models has been noticed, and it is worth referring here to Richard Levins' account of trade-offs between the desired qualities of models in population biology:

[28] Chirimuuta (2014; 2017; 2018) makes the case that "efficient coding explanations" offer answers to this kind of question. 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."

> It is of course desirable to work with manageable models which maximize generality, realism, and precision toward the overlapping but not identical goals of understanding, predicting, and modifying nature. But this cannot be done. (Levins 1966: 421)

We can add *intelligibility* to Levins' list of desiderata and note that the early generations of encoding models achieved precision (i.e. making quantitatively precise predictions) and intelligibility, but lacked realism both in the sense of goodness/poorness of fit to observed neural data and in terms of the accuracy of assumptions made about the neural systems. An artificial network used to model neural data can achieve unprecedented realism in the first sense, while the notion of realism of assumptions does not apply when its design is hypothesis free (as in the examples given above). However, there is a third relevant sense of realism – that of accurately capturing the computation performed by the biological neurons. Even without the function learned by the ANN being made explicit, it can safely be assumed that the models are more realistic in this sense, and that this is the source of their increased predictive accuracy. It is also significant that Levins distinguishes between the three goals of research – understanding, predicting, and modifying nature – while presupposing that if you aim well at one of these you are virtually guaranteed to hit the others. In my cases these goals also come apart: by aiming at prediction one cannot thereby expect to arrive also at understanding.

2.4 Opening the Black Box?

One obvious question about the trade-off outlined above is whether 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 simulations, such models will become as intelligible as the traditional sort. As mentioned above, the criterion of intelligibility is a context-sensitive one, so my analysis 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. I will argue in this section that nonetheless the ANN's will remain relatively less intelligible than their hand-coded counterparts, and that the trade-off will persist.

"Explainable AI" (XAI) is currently an active area of research.[29] But the focus is on the development of networks which self-report the basis for their classifications

[29] https://www.darpa.mil/program/explainable-artificial-intelligence

and is not of much benefit in theoretical neuroscience, precisely because of the lack of an established theory which could be used to assess the appropriateness of the AI's self-reported decision process. In the case of RNN's, methods have been developed to reverse-engineer trained networks in order to understand what features of the dynamical system are responsible for it arriving at the solution to the task.[30] Advocates of such methods argue that even though the artificial network is not well understood, it is still "vastly more accessible for research" than the actual brain (Barak 2017:3; cf. Gao and Ganguli 2015:151; McIntosh et al 2017:7-8; Cadena et al. 2017:13). Furthermore, 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.

My discussion proceeded above as if intelligibility were an all or nothing quality. This was an over-simplification because models can be relatively more or less intelligible – we should imagine there is a full greyscale range of boxes. For example, a scaled down version of a CNN would be less powerful but more intelligible than the full model; familiarising oneself with the properties of the toy model could help the full-scale model become somewhat more transparent to the user.[31] Given the context dependency of intelligibility, it could well be that as the process of reverse engineering the most successful AI models of the brain advances, scientists come to learn a new set of concepts for classifying neural activity, concepts that first originate from examining the network's behaviour.[32] Such an extension of neuroscientists' conceptual repertoire could facilitate the qualitative understanding of their models that is required by the criterion of intelligibility. Of course it remains to be seen whether AI will bear this sort of conceptual fruit.

Work on visualisation of information processing within CNN's is ongoing. In particular, the response properties of nodes in deep networks trained for object recognition have been mapped to reveal a hierarchical structure of artificial "receptive fields" going from simple to complex response properties, not unlike the mammalian visual system (Yamins and DiCarlo 2016). Since visualisability is often a hallmark of intelligible theories and models, this is highly relevant to my

[30] Sussillo and Barak (2013); for discussion see Chirimuuta (2017: §4).

[31] I thank Andy Clark for this suggestion. See Beer and Williams (2015) on toy models.

[32] This would be analogous to the eliminativist's prediction that neuroscientific concepts will replace folk psychological ones; here the AI-derived concepts would replace classic neuroscientific ones. I'm grateful to Suilin Lavelle for this point.

argument. One limitation of the kind of visualisations on offer here is that they do not come with the benefits of the qualitative pictures associated with the traditional models mentioned above. The deep network visualisations reveal something about information processing in the artificial network, but should not also be taken as representations of the biological network.

It should also be emphasised that such visualisations fall 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 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. We can say already that those functions must be nonlinear, and contain very many more terms than the ones found in the hand-written 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 (the criterion of intelligibility), as is possible with the very simple equations at the heart of the linear encoding models. 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 far less intelligible than a hand-coded model and for this reason the trade-off between prediction and understanding will not go away.

3. IS UNDERSTANDING FACTIVE?

To recap, a longstanding goal of neuroscience has been to produce predictively accurate models of neural responses to 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 capture accurately the complex nonlinear functions computed by real 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 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 true beliefs (Khalifa 2017:155). This indicates that a *non-factive* account of understanding[33] is required to make sense of my case studies and the finding of the trade-off so in this section I offer an exposition and defence of the non-factivist viewpoint.

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.[34]

However, there are important arguments against this proposal that understanding is non-factive because it is conferred by idealised models (Khalifa 2017: chapter 6). In particular, I will consider the objections that the understanding conferred by the traditional models is factive[35] because those models are (1) *approximately true* and (2) that the models offer understanding in virtue of their containing a *kernel of truth* in amongst the false assumptions. It turns out that modification to an account of abstraction and idealisation that has been endorsed by some non-factivists is needed in order to meet these objections.

Angela Potochnik presents a compelling picture of causal complexity as the reason for the prevalence of idealised models in science: observable nature

[33] I.e. one that denies that understanding requires belief in true or approximately true explanations of the phenomenon (Khalifa 2017:156).

[34] "[A]n 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).

[35] Or *quasi-factive* (Khalifa 2017:154-5). For simplicity, I present the issue as a disagreement between factivists and non-factivists.

presents us with countless regularities, some of which are relevant to human interests, and some are more reproducible than others. For scientific representations to be useful for understanding and action, they must be judicious in their choice of what to represent, and in their mode of representation. According to Potochnik causal patterns are genuine regularities in natural phenomena, though they may not be apparent -- and understandable -- without the aid of idealised model:

> Phenomena embody lots of causal patterns; grasping any old causal pattern embodied by some phenomenon won't lead to an understanding of that phenomenon. The grasped causal pattern must relate in the right way to the inquiry for it to produce understanding. (Potochnik 2017:116)

Like Elgin (2004), Potochnik draws on Dennett's notion of a *real pattern* to convey the idea of causal complexity and the benefit conferred by an idealised model (Dennett 1991: figure 1; figure 1 below). The real pattern "bar code" (alternating solid blocks of white and black pixels) is instantiated in each part of Figure 1, with different levels of noise overlaid. Potochnik suggests that the role of the idealised model is to isolate the pattern from the background of noise, and that this has practical advantages as well as the epistemic benefit of understanding.

> I have suggested that idealizations contribute to understanding by representing as-if to the end of depicting a causal pattern, thereby highlighting certain aspects of that phenomenon (to the exclusion of others) and revealing connections with other, possibly disparate phenomena that embody the same pattern or, in some cases, that are exceptions to that pattern. Potochnik (2017:97)[36]

In short, the observable phenomena comprise the real pattern (of interest) mixed with "noise" (which may turn out to be other causal patterns, not currently of interest). The job of the idealised model is to separate the wheat from the chaff, making the real pattern of interest more salient, comprehensible and useful.

[36] Cf. Elgin (2004:127): "If, e.g., we draw a smooth curve that skirts the data, and construe the data as a complex of relevant and irrelevant factors (signal and noise), …., we impose an order on things, highlight certain aspects of the phenomena, reveal connections, patterns and discrepancies, and make possible insights that we could not otherwise obtain (Dennett 1991). We put ourselves in a position to see affinities between disparate occurrences by recognizing them as variations on a common theme."

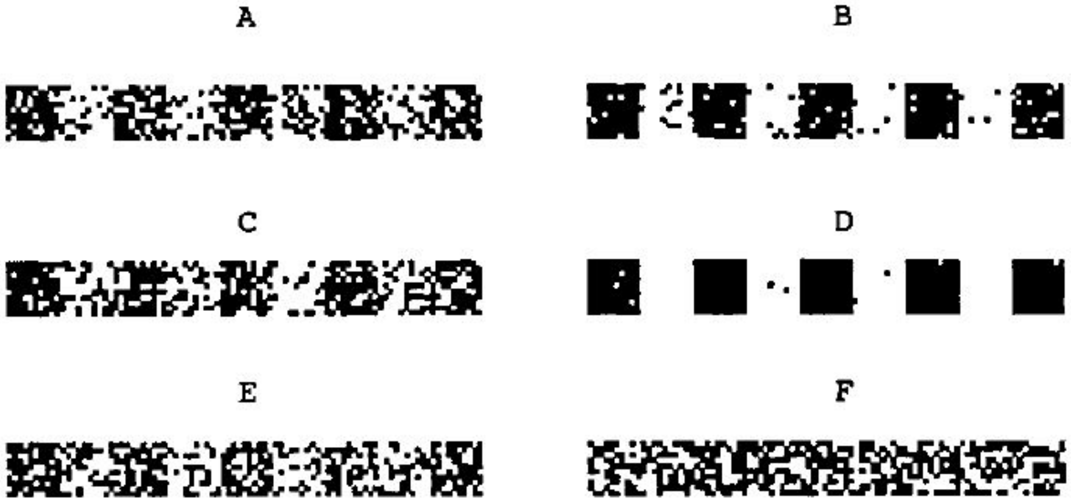


Figure 1 Examples of the "real pattern" bar code. The pattern (most clearly visible in D) is an alternating series of solid black and white squares. The pattern is obscured by varying amounts of noise. From Dennett (1991, figure 1) [permission needed].

In order to apply Khalifa's objections to non-factivism, I will follow him in assuming that non-factivist understanding is associated with false explanations (Khalifa 2017:156). In the case of the earlier generation neural models, I focus on the explanation of "what is computed?" – i.e. the statement of the linear function (or linear function with nonlinear bells and whistles) which is purported to be the computation performed by the neurons.[37] The greater predictive success of the ANN models is good grounds to say that those earlier explanations are false because the computations performed by these neurons are not, essentially, linear summations.

Against the non-factivist claiming that models providing false explanations nonetheless confer understanding, Khalifa (2017:162-4) offers the response that such explanations may be approximately true. Thus, one could say that the traditional models are the linear approximations of the true nonlinear functions computed by the neurons. This is a tempting response, but my concern is that it introduces a far too liberal standard for approximate truth. Any nonlinear function can be approximated by a linear one, but that does not make the linear function also *approximately true.* There would have to be some agreed empirical standard

[37] I restrict attention here to these explanations because we have greater clarity regarding their truth or falsehood. The efficient coding explanations associated with these linear models *might* still be true even if the functions are a false representation of the nonlinear neural computations.

of what counts as an acceptable margin of error, within which the predictions of the linear approximation must lie in order for that function to count as approximately true. The fact that the neuroscientists themselves have not considered the linear models satisfactory for their own predictive purposes suggests that such models would not satisfy an agreed standard.

However, I believe that problems do arrive for the non-factivist if too much is made of the "real pattern" account of the relationship between the phenomenon in nature and the idealised and/or abstract model. For example, if one takes "bar code" to be "really there" in the natural phenomena, but only masked by different degrees of noise and causal complexity, then it does seem natural to say that a hypothetical model which represents any of the patterns depicted in Figure 1 A-F as "pure bar code" (solid alternating squares of black and white) is approximately true even if the error rate is extremely high (as it is for E and F).

Furthermore, the real pattern account opens the non-factivist to another line of attack -- the idea that the false explanation contains a kernel of truth, and it is the true (i.e. genuinely referring) component of the explanation that confers understanding (Khalifa 2017:173-5). If one conceives of the phenomenon in nature as the sum of a causal pattern of interest, plus other complex causal patterns, plus random noise, then it is natural to say that the true kernel of the model accurately represents the causal pattern of interest, while the abstract and/or idealised elements of the model are psychological aids to the scientist (Sullivan and Khalifa under review) or ways of flagging irrelevant causal factors (Strevens 2008). For instance, the true kernel of a hypothetical model could accurately represent bar code, while the remaining elements inaccurately depict (by omission or distortion) the more complex features of the phenomena -- the ones due to noise or other complex causal patterns. In the case of the neural models, one could assert that the linear response pattern is a *real pattern*, manifest in certain experimental conditions (e.g. when the visual system is simulated with very simple artificial images), and that the traditional models accurately represent that pattern while inaccurately representing (by omission or distortion) the non-linear behaviour of the cell. If this is the case, then the explanation offered by the linear model would *not* be false, and understanding would turn out to be factive after all.

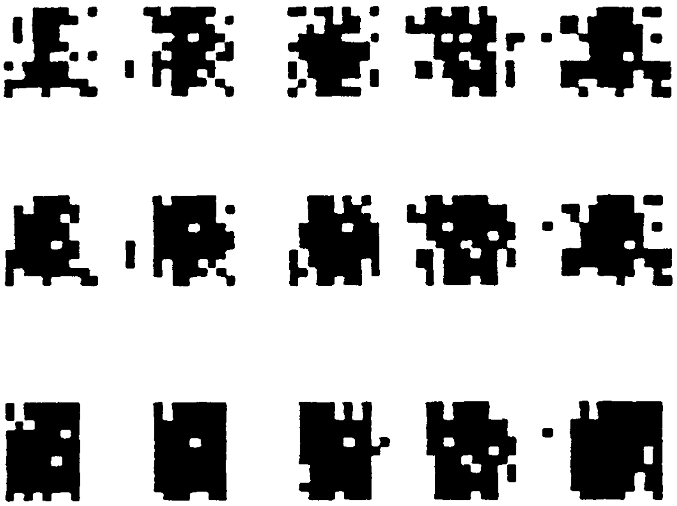


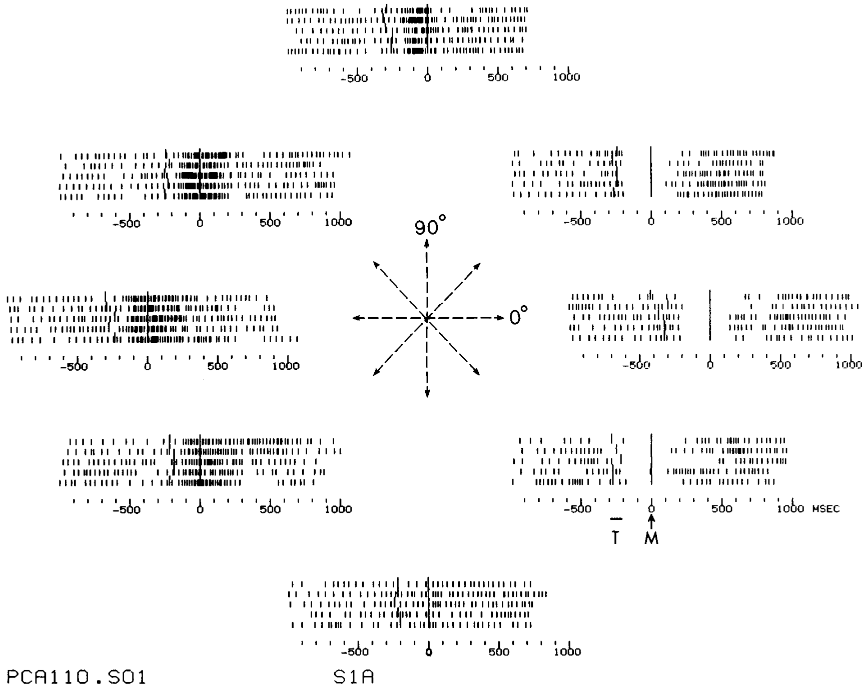
Figure 2. An "Ideal Pattern". The top row is generated by a stochastic process. Pixels are drawn from normal distributions, evenly spaced along the row. Hard edges appear in the second and third row through application of a contrast enhancer – such results of averaging are what I call "ideal patterns". From Dennett (1991, figure 4). [permission needed]

In order to avoid this result, the non-factivist should appeal to another kind of pattern depicted by Dennett (1991, figure 4; see figure 2 above), which I will call an "ideal pattern".[38] Such patterns do not arise from the superposition of a simple pattern with noise and other complex regularities. Instead, the original pattern (top row) is highly irregular, unlike bar code, but is made more regular though application of a filter (second and third rows). As Dennett (1991:44) describes,

> The frames in figure 1 were created by a hard-edged process (ten black, ten white, ten black, . . .) obscured by noise, while the frames in figure 4 [figure 2 above] were created by a process almost the reverse of that: the top frame shows a pattern created by a normal distribution of black dots around means at x = 10, 30,50, 70, and 90 …; the middle and bottom frames were created by successive applications of a very simple contrast enhancer applied to the top frame: a vertical slit "window" three pixels high is thrown randomly onto the frame; the pixels in the window vote, and majority rules. This gradually removes the salt from the pepper and the pepper from the salt, creating "artifact" edges such as those discernible in the bottom frame.

[38] For Dennett (1991) both sorts of patterns were real in a mild sense -- *quasi-real*.

In order to employ this notion of pattern to our case, I will draw an analogy between the contrast enhancer and the data processing operations that are ubiquitous in science, and which generate the *phenomena* -- in the sense of Bogen and Woodward (1988) -- that are the actual target of modelling and theory building, as opposed to the raw data. A process as simple as averaging across trials creates "hard edges" in the phenomena that are not there in the raw data. For instance, neural recordings show trial to trial variability regarding the exact onset time and rate of activity in response to a stimulus (see figure 3a). Traditional neural models target average data – the simplified and regularised patterns that are created through data processing. In this sense, the target of the model is an "ideal pattern" (a pattern that is partially human dependent) rather than a "real pattern" (a pattern that is just out there in nature, waiting to be revealed).



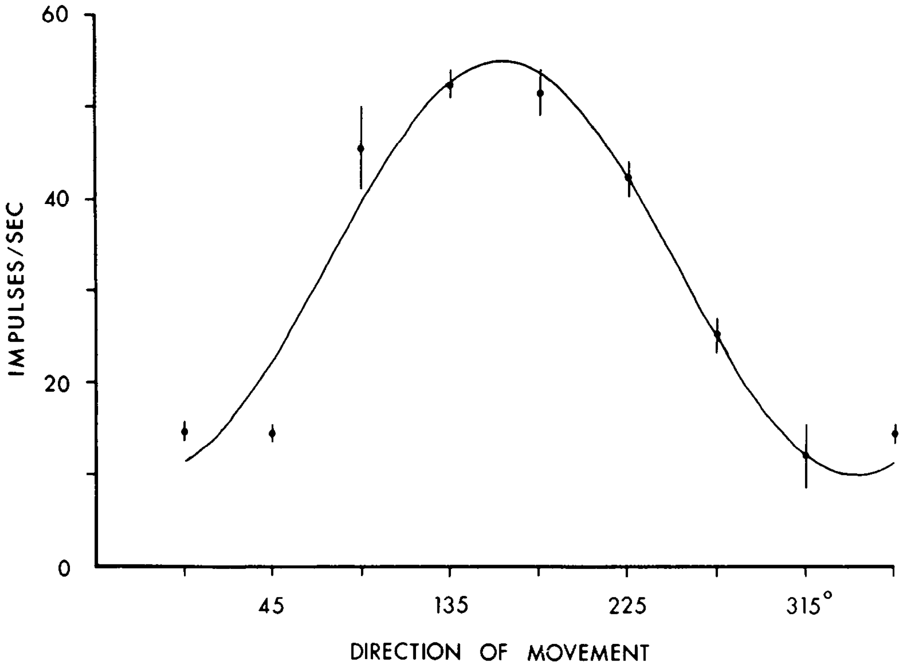


Figure 3a – Raster plot. Spike data for five trials, for one neuron responding to eight different directions of movment.
Figure 3b – The Cosine model of a neuron's tuning to direction of movement. The cosine function is fit to trial average data. From Georgopoulos et al. (1982) [permission needed]

To illustrate, Figure 3a shows the spike data for five trials of one neuron's responses to eight directions of movement. Onset time and level of excitation (or inactivation) in response to movements vary across trials but the pattern that is the actual target of modelling is a trial averaged one which is, strictly speaking, a creation of the scientist. The brain itself does not average across trials and each instance of variation in trials could be "meaningful" to the brain – reflecting different modes of attention or posture – rather than pure noise. On Potochnik's picture, the target of the idealised model was the "real pattern" in nature. Instead, we now have a four-place scheme of model, phenomenon, raw data and natural events. The model target is the "ideal pattern" – the phenomenon created by processing the raw data, where the raw data more closely track the natural events. On this picture, the opponent to non-factivism cannot simply say that some components of the model do accurately represent (or approximate to) the natural events, because the specific targets of the model are not natural events per se but the phenomena – the "hard edges" introduced by the scientists' processing of data.

I will happily admit that there is a close connection – in visualisable cases, typically a resemblance – between the events in nature and the phenomena targeted by the model. For example, all the rows of figure 2 show periodic increases and decreases in pixel density. In figure 3, the raw spike data (a) and the averages (b) both show a systematic increase and decrease in activity levels. It is because such relationships obtain that idealised models are useful. However, the features represented specifically by the model – the hard edges of figure 2, and in our case studies, the linearity of neuronal responses – are not, strictly speaking, there in the natural events. So my contention is that the linearity represented in the traditional neural models should be thought of as an "ideal pattern" -- not actually there in the brain but projected onto it in the course of an interactive process in which the scientific activities of data gathering and processing grapple with the complexity of the brain.

In sum, I have argued in this section that the trade-off between prediction and understanding can best be accounted for a non-factivist account of scientific understanding,[39] and that non-factivism is most tenable if we construe the intelligible models of neural activity as representing ideal rather than real patterns. One might be wondering, *if understanding is brought about, in such cases, by the modelling of patterns not simply there in nature, why should we want it? Should we not resist this version of understanding that comes to us via human-generated "illusions"?* This worry is misguided just because the "ideal pattern" is not a mere illusion, something projected onto the natural system and unconstrained by the natural events. As stated above, it is the result of an interactive experimental process and is therefore quite tightly constrained. The point of calling it an "ideal pattern" is just to highlight the fact that it is not entirely human-independent; but it is not entirely nature-independent either – it is both scientist and nature dependent. As stated above, the intuition behind the non-factivist account is that understanding is a compromise between the recognition of complexity in nature and the limitations of humans' ability to represent complex events. The notion of an "ideal pattern" is here intended to encapsulate this compromise.

[39] To reinforce this point: if the factivist is correct and understanding is a product of obtaining true scientific beliefs, then knowing the true function computed by the neurons should bring about more understanding. I argued in Section 2.4 that this would not be the case, even if the function learned by the ANN was made explicit to the scientist. In other words, the divergence of accurate prediction and understanding is the opposite of what a factivist account of understanding would predict.

Khalifa (2017:178) proposes one final move for blocking the non-factivist account. The idea is that the scientist's *acceptance* of a false explanation (the appropriate epistemic attitude, instead of belief) is a kind of scientific knowledge, alongside belief; thus understanding is knowledge dependent, and therefore factive, after all. This liberal attitude to scientific knowledge is actually quite conducive to the non-factivist account, as I have presented it, because it dislodges the notion that the apprehension of the True Beliefs about nature is the one way to have scientific knowledge. It captures the idea that in accepting explanations founded on models depicting ideal rather than real patterns, we cannot claim that all our understanding and scientific knowledge is based on the apprehension of transcendent Truths about human-independent nature. That is a modest and reasonable picture of scientific knowledge.

4. CONCLUSION

One appeal of using advanced AI models in science is that they make possible unprecedented degrees of predictive accuracy, and therefore technological control. I have argued that such benefits are likely to come at a price – that of understanding the system that is being modelled and manipulated. I have argued 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 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 forced on us by the trade-off outlined in this paper.

References

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

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

Bogen, J., and J.F. Woodward. 1988. 'Saving the Phenomena', *The Philosophical Review*, 97: 303-52.

Cadena, Santiago A., George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker. 2017. "Deep convolutional models improve predictions of macaque V1 responses to natural images." In *bioRxiv*.

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.

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.

———. 2017. 'Explanation in Computational Neuroscience: Causal and Non-causal', *The British Journal for the Philosophy of Science*.

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

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.

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.

Dennett, Daniel, C. 1991. 'Real Patterns', *Journal of Philosophy*, 88: 27-51.

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.

Fan, Joline M., Paul Nuyujukian, Jonathan C. Kao, Cynthia A. Chestek, Stephen I. Ryu, and Krishna V. Shenoy. 2014. 'Intention estimation in brain–machine interfaces', *Journal of Neural Engineering*, 11.

Finkelstein, G. 2013. *Emil du Bois-Reymond: neuroscience, self, and society in nineteenth-century Germany* (MIT Press: Cambridge, MA).

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.

Georgopoulos, A.P., J.F. Kalaska, R. Caminiti, and J.T. Massey. 1982. 'On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex', *The Journal of Neuroscience*, 2: 1527-37.

Gilja, V., P. Nuyujukian, C. A. Chestek, J. P. Cunningham, B. M. Yu, J. M. Fan, M. M. Churchland, M. T. Kaufman, J. C. Kao, S. I. Ryu, and K. V. Shenoy. 2012. 'A high-performance neural prosthesis enabled by control algorithm design', *Nat Neurosci*, 15: 1752-7.

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.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. 'Deep learning', *Nature*, 521: 436-44.

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, Lane T., Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen A. Baccus. 2017. "Deep Learning Models of the Retinal Response to Natural Scenes." In *arXiv:1702.01825*.

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).

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

Stanley, G. B. 2013. 'Reading and writing the neural code', *Nat Neurosci*, 16: 259-63.

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

Strevens, Michael. 2008. *Depth: an account of scientific explanation* (Harvard University Press: Cambridge, MA).

Sullivan, Emily, and Kareem Khalifa. under review. 'Idealizations and Understanding: Much Ado About Nothing?'.

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.

Sussillo, D., S. D. Stavisky, J. C. Kao, S. I. Ryu, and K. V. Shenoy. 2016. 'Making brain-machine interfaces robust to future neural variability', *Nat Commun*, 7: 13749.

VanRullen, R. 2017. 'Perception Science in the Age of Deep Neural Networks', *Front Psychol*, 8: 142.

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.