Don't Fear the Bogeyman: On Why There is No Prediction-Understanding Trade-Off for Deep Learning in Neuroscience

Barnaby Crook & Lena Kästner

University of Bayreuth, Germany

barnaby.crook@uni-bayreuth.de / lena.kaestner@uni-bayreuth.de

Abstract

Machine learning models, particularly deep artificial neural networks (ANNs), are becoming increasingly influential in modern neuroscience. These models are often complex and opaque, leading some to worry that, by utilizing ANNs, neuroscientists are trading one black box for another. On this view, despite increased predictive power, ANNs effectively hinder our scientific understanding of the brain. We think these worries are unfounded. While ANNs are difficult to understand, there is no fundamental trade-off between the predictive success of a model and how much understanding it can confer. Thus, utilizing complex computational models in neuroscience will not generally inhibit our ability to understand the (human) brain. Rather, we believe, deep learning is best conceived as offering a novel and unique epistemic perspective for neuroscience. As such, it affords insights into the operation of complex systems that are otherwise unavailable. Integrating these insights with those generated by traditional neuroscience methodologies bears the potential to propel the field forward.

Keywords: deep learning, neuroscience, machine learning, trade-off, accuracy, prediction, scientific understanding, explainability, epistemic perspective

1. Introduction

Machine learning in general, and deep artificial neural networks (ANNs) in particular, have become an increasingly influential methodological tool in the arsenal of modern computational neuroscience. Because of their flexibility, ANNs are used in multiple distinct ways, from the identification of predictive variables to serving as models of cognitive and perceptual faculties (Glaser et al., 2019). As a result of their expressive power, ANNs have achieved unmatched success when it comes to task performance, prediction of neural activity, and replication of human-like behavioural patterns (Cao & Yamins, 2021; Kanwisher et al., 2023; Schrimpf et al., 2020). However, despite their undeniable predictive power, deep learning-based approaches to investigating the brain are not without limitations. One such limitation is that systems designed through deep learning are often highly complex and opaque to researchers. Because their functional structure is generated through an automated learning procedure, precisely *how* deep learning-based systems solve a given set of computational problems often remains unclear, even to their designers. This applies not only to the use of deep learning in neuroscience but to deep learning systems at large. Many AI researchers worry that there is a "tension between machine learning performance (predictive accuracy) and explainability" (Gunning & Aha, 2019, p. 45; see Crook et al., 2023 for discussion). This general worry is also echoed with respect to neuroscience research specifically. For instance, Lindsay argues that "to have a model complex enough to perform real-world tasks, we must sacrifice the desire to make simple statements about how each stage of it works" (2021, p. 2024). Building on such worries, Chirimuuta portrays the situation as one where computational neuroscience faces "a trade-off between predictive accuracy and the ability of [its] models to confer understanding" (2021, p. 787).

While we agree with Lindsay's statement, we believe that the fear expressed in Chirimuuta's is misguided. There might indeed be a trade-off between the predictive success of individual models and the ease of understanding them; but this does *not* imply that utilizing complex computational models in neuroscience will generally inhibit our ability to understand the (human) brain. On the contrary: We firmly believe that rather than worrying about sacrificing understanding by utilizing deep learning in neuroscience, we should welcome ANNs into the neuroscientists' toolkit. They bear the potential to generate new insights about the brain from a yet unavailable *epistemic perspective* (see Kästner, 2018) that is complementary to the

perspectives provided by established neuroscientific methodologies. That is, deep learning contributes to, and does not detract from understanding the phenomena neuroscientists investigate.

We shall begin our exposition by diagnosing the source of trade-off worries in a misconception of scientific understanding (section 2). The reasoning behind the notion that researchers are paying a high price for predictive success crucially depends on weighing the predictive power of models against the intelligibility of those models. While it is widely agreed that the aim of science is to understand *phenomena*,[1] the argument for a trade-off depends on equating this understanding with the intelligibility of particular models of those phenomena (see, e.g., Bokulich, 2017; Elgin, 2017).[2] However, we think this is mistaken in two ways. First, while we agree that models play a key role in the production of scientific understanding, we think such understanding depends upon integration of numerous distributed sources of explanatory information. Second, the intelligibility of a model is not a reliable guide for how much understanding it can confer. While interpreting complex models may be challenging, so long as it is feasible, even relatively unintelligible models can induce scientific understanding of phenomena.

Once we recognize this distinction between model understanding and *model-induced understanding of phenomena*, it becomes evident that the prediction-understanding trade-off is a bogeyman we need not fear: Deep learning can contribute to a progressive computational neuroscience that yields both improved predictive performance *and* deeper scientific understanding (section 3). This can be achieved by integrating insights from traditional neuroscientific research with those delivered by ANNs and systematic post-hoc interpretability strategies. Insights from these different *epistemic perspectives* will contribute to a store of distributed knowledge that neuroscientists can draw on to explain and understand phenomena.

Because debates about what it takes to explain and understand complex systems have a particularly rich history in neuroscience, we focus our discussion on this field. Still, we think

[1] There is a rich debate in contemporary philosophy of science about what precisely phenomena are and how to best characterize them (e.g., Colaco, 2019; Craver & Darden, 2013; Feest, 2017; Shagrir & Bechtel, 2017). Here we simply use the term "phenomenon" to refer to any explanandum investigated by neuroscientists. Prototypical examples include such things as color vision, face recognition, or cortical language processing.

[2] We use the term *intelligible* to capture how easy it is to understand a given model.

our insights will not only apply to neuroscience; but are likely to generalize to relevantly similar domains where deep learning is being utilized to study complex systems.

2. Trade-Off Worries

In this section, we discuss why trade-off worries come about. We shall first review arguments to the effect that a trade-off exists between the predictive success of a model and how much understanding it can confer (section 2.1). Against this backdrop, we shall explain why we think these arguments are flawed: because they mistake model understanding for model-induced understanding of phenomena (section 2.2).

2.1 Why worry?

A common conception is that models inducing understanding should be simple and intelligible, abstracting away from or usefully idealizing the complexities of the real world (cf., Bokulich, 2017; Elgin, 2017; Potochnik, 2016). Against this backdrop, it seems natural to expect that scientific understanding is best conferred by simple transparent models that domain experts understand perfectly. In traditional scientific modelling, scientists' domain knowledge is encoded by the construction of a model of a target phenomenon (Craver, 2001; Frigg & Hartmann, 2020). These models can frequently be represented by intuitive box-and-arrow diagrams or by simple equations experts can grasp just by eyeballing them. The components of these models are often taken to represent relevant features of target phenomena, though these may be individuated functionally rather than structurally (Bennett et al., 2019; Craver & Kaplan, 2020). The construction, evaluation, and manipulation of models that represent phenomena is often taken to be central to the production of scientific understanding (de Regt, 2009; Elgin, 2007; Morgan & Morrison, 1999).

Deep learning models are different on various scores.[3] First, because they are trained with large datasets and consist of many parameters, they can be exceedingly complex. As such,

[3] The word 'model' has multiple meanings. In statistics, any deep learning system is called a model, regardless of the purpose for which it is constructed. However, in the philosophy of science, a model is any structure, mathematical, physical, or graphical, that is constructed to represent (some aspect of) a target system. In neuroscience, statistical models based on deep learning are used both as tools (e.g., for data pre-processing; c.f. Glaser et al., 2019) *and* as models that represent target systems (Cichy & Kaiser, 2019). In what follows, we focus on the use of ANNs as *scientific* models (i.e., models constructed to represent), which is of greater philosophical interest.

they break with the expectation of being simple. Second, and relatedly, unlike with traditional modelling, domain knowledge typically only constrains choices about the initial *architecture* of deep learning-based models (Goodfellow et al., 2016). This consists of the number of layers, number and types of neurons in each layer, connectivity profiles of each neuron, and learning rules of the network (Chollet, 2021; Richards et al., 2019). While this architecture is specified by researchers, the rest of what an ANN model encodes (e.g., the system's functional units, their organisation, and how they interact to solve the computational problem) is learned through an automated training procedure, typically involving backpropagation and some form of gradient descent (LeCun et al., 2015). On the one hand, this autonomy from human design is the central strength of deep learning. It is what gives ANNs the flexibility to learn subtle, task-relevant structure that eludes traditional modelling and thus to expand the range of domains in which models can achieve predictive success (Boge, 2022; Jumper et al., 2021; Zhuang et al., 2021). On the other hand, especially when ANNs have many parameters, it is often unclear both i) how the internal structure of the model implements the mapping from input to output (this is 'w-opacity' in the terminology of Boge, 2022; see also Chirimuuta, 2021; Creel, 2020), and ii) what relation this internal structure bears to the target phenomenon (this is what Sullivan, 2022 calls 'link uncertainty').[4] As such, deep learning models can be predictively successful without necessarily illuminating how exactly the target phenomenon is produced; that is, without conferring understanding.

As a result of these unique properties, several authors writing about the increasing prevalence of deep learning in science have worried that the advantages we accrue in terms of predictive accuracy are offset by a decrease in explanatory power and understanding (Boon, 2020; Chirimuuta, 2021; Lindsay, 2021; López-Rubio & Ratti, 2021; Srećković et al., 2022). For example, Srećković and colleagues suggest that increased use of deep learning models may "lead us away from the explanatory aspect of science" (2022, p. 172). Along similar lines, Boon imagines a future in which "scientific researchers and scientific knowledge become superfluous as learning from large data sets, algorithms and data-models will be developed at a degree of complexity and adequacy far beyond the capacity of the human intellect" (2020, pp. 49–50). Finally, López-Rubio and Ratti express the view thusly: "when predictive

[4] There is a vast body of literature on opacity. For a classic see Burrell (2016), for a recent comprehensive taxonomy see Mann et al. (2023).

performances increase, the possibility of elaborating a (mechanistic) explanation necessarily decreases" (2021, p. 1332). While the exact claims and arguments of these authors differ slightly, the overall perspective they present is clear: we must choose between increasing the predictive adequacy of our models and gaining understanding from them; we cannot have both.

Worries about this hypothesized trade-off between predictive success and understanding have touched many areas of science, from protein-folding (Bouatta et al., 2021), to physics (Boge, 2022), to molecular biology (López-Rubio & Ratti, 2021). In what follows we concentrate on neuroscience, the discipline with which we are most familiar. Though we spell out our response with this specific focus, we suggest that the broad shape of our argument is likely to carry over to relevantly similar scientific domains, even if the details may differ.

Perhaps due to the intertwined history of neuroscience and AI research (Hassabis et al., 2017; McCulloch & Pitts, 1943), deep learning has had a particularly large and rapid influence on the study of the brain and nervous system (Cichy & Kaiser, 2019; Marblestone et al., 2016; Richards et al., 2019). As well as being used to identify predictive features and solve engineering problems (Glaser et al., 2019), deep learning has been employed in computational neuroscience to produce ANN models of cognitive and perceptual systems (Doerig et al., 2023; Kell et al., 2018; Schrimpf et al., 2020; Yamins & DiCarlo, 2016). It is in this capacity that deep learning's contributions to neuroscience have prompted worries of a trade-off. While ANNs have proven successful at implementing solutions to ethologically relevant tasks, predicting neural activity, and recapitulating human-like error profiles, critics have raised concerns that we are simply replacing one black box with another (Charles Leek et al., 2022; Chirimuuta, 2021; Thompson, 2021). This argument has been made explicitly by Mazviita Chirimuuta, who claims that computational neuroscience faces "a trade-off between predictive accuracy and the ability of [its] models to confer understanding" (2021, p. 787). We shall briefly reiterate her argument before explaining why we do not agree with its conclusion.

Chirimuuta directs her argument at models "intended to represent the functions computed by neural systems" (2021, p. 772).[5] As she details, traditional models, such as Georgopoulos

[5] Chirimuuta argues for a principled distinction between *efficient coding* models which aim to capture the encoding functions used by the brain and *mechanistic* models which (also) encode anatomical isomorphisms, intending her argument to apply only to the former. However, in our view, the properties of ANNs that Chirimuuta

and colleagues' (1986) *Population Vector Algorithm* for estimating arm movements from primary motor cortex activity, are fairly simple. Although such models knowingly encode false assumptions, in Chirimuuta's estimation they end up being "highly intelligible, representationally inaccurate but surprisingly useful" (2021, p. 774). At the other end of the spectrum, models based on deep learning, such as Sussillo and colleagues' (2012) recurrent neural network for decoding arm kinematics from neural activity, are complicated beasts. As Chirimuuta puts it, such models are "non-linear, opaque but predictively accurate" (2021, p. 776). This contrast recapitulates the distinction between traditional modelling strategies and deep learning that we have already discussed, leading to the suggestion that "there is a trade-off between a model's predictive power and its ability to increase the scientist's understanding of a neural response" (Chirimuuta, 2021, p. 768). However, we think this is too quick. To reach this conclusion, Chirimuuta uses the notion of 'model intelligibility' to capture not just the idea that models are easy to understand but *also* their overall ability to confer understanding. It is the conflation of these two notions that we take issue with.[6]

At this point, it is important to introduce a subtle but important distinction concerning the source of the unintelligibility or *opacity* of ANNs.[7] On the one hand, ANNs are trained via automated learning procedures. This is what makes it possible for researchers to develop models that solve certain tasks without understanding exactly how they do so. On the other hand, ANNs are typically very complex, consisting of many tuneable parameters which interact in subtle ways to produce the behaviour of the whole system. While these two sources of opacity are intimately related in the methodology of deep learning, they are conceptually distinct. The first problem is that an explanation is not *given* (i.e., automatically conferred through construction of the model), the second is that an explanation is not *simple*. In principle, the first source of opacity might be overcome through systematic post-hoc analysis and interpretation of a trained ANN (as we will discuss in section 3). Assuming we might

invokes to make her case creates a structurally analogous problem for both kinds of model. As such, we take our response to address a generalised version of Chirimuuta's argument. This argument covers a broad class of ANN models in neuroscience, including those with mechanistic explanatory goals.

[6] We wish to stress that this is not a semantic dispute about how *intelligibility* should be defined, but a conceptual dispute about whether 'ease of understanding' and 'ability to confer understanding' ought to be equated with one another.

[7] We take opacity to mean "barrier to conferring understanding". Notice that opacity is at least partially a subjective phenomenon. What is opaque to a layperson may not be opaque to an expert (Humphreys, 2009).

reverse engineer the functional structure of ANNs, however, we are still left with the second source of opacity, viz. the structure we uncover might be quite complex. Chirimuuta states clearly that in her view this second source is sufficient to ensure the trade-off obtains:

> "[…] even if we could write down by hand the equations embedded in the trained ANN's […], 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." (2021, p. 782)

So, even if an explanation was given, it would not necessarily alleviate opacity because it might lack the simplicity required for humans to understand what a system is doing. In other words, even sophisticated interpretability methods may be powerless to overcome the trade-off between the predictive success of deep learning models and their potential to confer understanding. With these arguments in view, let us turn to our own assessment of the trade-off story.

2.2 Why not to worry!

We grant that there might be a trade-off between predictive success and ease of understanding for individual models. However, we do not think that this prevents neuroscientists from gaining understanding by utilizing ANNs in their research. This is because what they are after, we think, is not intuitive understanding of models of complex systems but understanding the *phenomena* that these systems exhibit with the help of models. Thus, there is no requirement for models to be simple. What matters is that they confer *understanding of the target phenomenon*, even if that understanding requires significant interpretative work. Indeed, as we will show below, there is nothing in principle stopping complex models from conferring greater understanding than simple ones.

Before we continue, a note on what we take (scientific) understanding to be is in order. Generally speaking, we are sympathetic to Elgin's suggestion that "understanding is a grasp of a comprehensive body of information that is grounded in fact, is duly responsive to evidence, and enables non-trivial inference, argument, and perhaps action regarding that subject the information pertains to" (2007, p. 39). Put simply, we take understanding to be having some

kind of *qualitative grasp* of a target phenomenon (de Regt, 2009; Elgin, 2007; Strevens, 2013), some kind of *insight* into how a phenomenon of interest is exhibited by a system. It is a widely acknowledged feature of grasping is that it is gradable, i.e., it comes in degrees (Elgin, 2007; Hills, 2016). With this in mind, we can say a subject S grasps a phenomenon P to the extent that they are able to (i) make coarse-grained predictions about P without performing exact calculations (cf., de Regt, 2009), (ii) give rich, detailed, and productive answers to a broad range of questions concerning P (cf., Chirimuuta, 2021), and (iii) intervene on and control P (or at least to describe such interventions, even if they would be impractical) (cf., de Regt, 2009). To serve these purposes, S might utilize some kind of (mental) representation of P where this representation might be considered an explanation for P or, alternatively, enable S to explain P. Given (i)-(iii) are also features commonly associated with good scientific explanations (see e.g., Craver, 2007; Pearl, 2009; Salmon, 1984; Woodward, 2003), this tight connection between understanding and explanation is no accident. If Alice has a better grasp of P than Bob (in the sense of (i)-(iii)), then she is likely to not only have a better understanding of P but to also be able to *explain* it better.

Admittedly, this brief characterization of understanding in terms of grasping is fairly minimal. What we say about the relation between (i), (ii), and (iii), about the role of (mental) representations, and about the precise relation between understanding and explanation remains intentionally sketchy. But since the argument we make in what follows is compatible with a range of more articulated accounts of understanding (see de Regt, 2009; Elgin, 2007; Hills, 2016; Strevens, 2013 for various detailed views; see Grimm, 2021 for an overview), we think this is a merit rather than a shortcoming. What is important for the current discussion is merely that scientific understanding is conferred by the very kind of information we utilize in scientific explanations. And that the targets of understanding and explanation in science are *phenomena* (e.g., Bechtel & Richardson, 2010; Colaco, 2019; Craver, 2007; Feest, 2017; Glennan, 2017; Hochstein, 2016; Kästner, 2018).

This takes us to our assessment of the trade-off arguments presented above (section 2.1). In those arguments, a trade-off between model complexity and intelligibility is taken to imply a trade-off between model complexity and the potential of a model to confer understanding of phenomena. But understanding models is not the same as understanding phenomena. And the ease of understanding a model is not the same as how much understanding that model

can confer. Granted, all of the tell-tale signs of grasping just described can be straightforwardly applied to models and their behaviour (i.e., by casting them as the target phenomenon). However, while models play a crucial instrumental role in our understanding of phenomena, they are not typically the ultimate *targets* of inquiry in neuroscience or any other natural science; models are usually mediators (c.f., Morgan & Morrison, 1999), not explananda. We must take care not to mistake model understanding for phenomenon understanding.[8]

The issue we are getting at is nicely captured by Lawler and Sullivan's (2021) distinction between *model explanation* and *model-induced explanation*.[9] In model explanation, the content of the model simply is the explanation. In model-induced explanation by contrast, "the relevant explanatory information is independent of the model but only closely intertwined with the model due to the history of obtaining the information" (Lawler & Sullivan, 2021, p. 1069). Talking about understanding we can reframe the point thusly: Whenever the content of our understanding of a phenomenon includes information extrinsic to any particular model, model-induced understanding of a phenomenon is not reducible to model understanding. We think this is plainly the typical case in scientific practice (cf., Hochstein, 2016; Kästner, 2018; Mitchell, 2002, 2019). As such, reducing understanding of phenomena to what is encoded in a model is both descriptively and normatively inadequate. It is descriptively inadequate because real-world scientific understanding of complex phenomena frequently depends upon facts and details extrinsic to any individual model. It is normatively inadequate because, unlike a more holistic view of understanding, it artificially prevents us from satisfying scientific desiderata that might otherwise be satisfied (i.e., increasing both predictive accuracy *and* understanding).

With these distinctions in mind let us return to Chirimuuta's argument. Recall her claim that "models of neural systems are either very intelligible, or predictively accurate, but not both"

[8] This is not to say model understanding cannot contribute to understanding phenomena. Indeed, model understanding plays a crucial role. However, the relationship between understanding models and understanding phenomena is complex. We will elaborate on this in section 3.2.

[9] Note that Lawler and Sullivan are talking about explanation rather than understanding. Though this kind of distinction is also utilized in the debate on understanding (Strevens, de Regt ???), Lawler and Sullivan make the distinction most succinctly. As we stated above, we see an intimate link between the two. Depending on how exactly the relation between understanding and explanation is being construed, the distinction might even collapses (e.g., when understanding is conceptualised as the possession of a mental representation of the right kind of explanation).

(2020, p. 781). In a certain sense, we agree with this claim. It is obviously *easier* to understand simple models than it is to understand complex models (perhaps tautologically so). However, that does not imply that we obtain a greater understanding of *phenomena* through constructing and using simple models than we do through constructing, using, and interpreting complex ones. In fact, complex models will often confer higher degrees of understanding of phenomena in the sense described above: They enable us to (i) make better predictions, (ii) answer more questions about a phenomenon, and (iii) allow for more intervention and control.

But what about the opacity of ANNs – does it not interfere with (i)-(iii)? To address this question, recall our distinction between whether explanatory information encoded in an ANN is *given* and whether it is *simple* (section 2.1). Suppose that the *givenness* issue is overcome and we obtain an equation describing the function the model has learned, e.g., through interpretability methods (we will address the feasibility of this in section 3). Chirimuuta argues that, even so, "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 [as traditional, simpler models]" (2021, p. 782). Here, we think Chirimuuta conflates two things: the ease of coming to understand something is not the same as the degree of understanding that one might obtain from it. To make the point clear, we can return to Linday's claim that "to have a model complex enough to perform real-world tasks, we must sacrifice the desire to make simple statements about how each stage of it works" (2021, p. 2024). Lindsay is probably right. However, we do not think that understanding of phenomena needs to consist in being able to make *simple* statements. Deep ANNs might be challenging to understand. It might be laborious to gain insights about how they work, those insights might not be simple, and establishing their relationship to target phenomena may be difficult. But they might still enable a deeper scientific understanding than highly intuitive but overly simple equations or box-and-arrow models. How easy it is to understand a model does not strictly determine how instrumentally valuable that model will be to understanding and explaining a phenomenon all things considered.

Importantly, partial understanding of a complex but accurate model can lead to greater understanding of a phenomenon than complete understanding of a simple but inaccurate model. To see this, let us focus on prediction for the moment (we will return to the other

criteria for understanding in Section 3). Consider the following toy example: a system S elicits a behaviour P (the phenomenon of interest). S consists of 5 interacting factors (f1, …, f5) and P is described by a single scalar value y that can be obtained by simple mathematical operations performed on values taken by f1, …, f5. Bob constructs a simple model to help him predict P. Bob's model consists of two predictors, x1 and x2, which could (but do not have to) approximate f1 and f2, respectively. To make predictions about P, Bob estimates x1 + x2 in his head, which he can do perfectly. As such, we can say that Bob understands his model perfectly; but (due to the simplicity of his model) his predictions about the phenomenon are very inaccurate. Meanwhile, Alice constructs a complex model to help her understand P. Alice's model consists of five predictors x'1, …, x'5 which approximate all the factors of the target system. To make predictions about P, Alice estimates 0.8*x'1 + 0.9*x'2 + 0.7*x'3 + 0.2*x'4 in her head. She usually does not get all the calculations right, sometimes mixes up the multipliers, and always omits x'5 completely.[10] Thus, we might say her grasp of the model is partial or incomplete, even though the model itself might be perfect. Still, Alice's predictions about P are reasonably accurate. So, despite Alice's model understanding being partial where Bob's is perfect, Alice's predictions about P are still better than Bob's. As such, all else being equal, we attribute greater *model-induced understanding* of P to Alice than to Bob.

This toy example makes it plain (again) that there is an important conceptual distinction between model understanding and model-induced understanding. Even if we can understand a simple model perfectly, we may gain less scientific understanding about a phenomenon P than we can from a complex model we only partially understand. However, whether we *actually* gain a deeper understanding of P by utilizing complex models remains an empirical question. Still, if our diagnosis is correct, then there is no principled trade-off between predictive accuracy and the potential for models to confer scientific understanding. There is no reason to fear the bogeyman! Utilizing deep learning in neuroscience need not systematically inhibit our understanding of the brain. Rather, we think, it can support our understanding of neuroscientific phenomena in previously unavailable ways. This takes us to the next section.

[10] This detail about x'5 mirrors cases where parts of the internal structure of ANNs remain opaque.

3. Understanding the Brain with Deep Learning

So far, we suggested that scientific understanding consists in having a qualitative grasp of target phenomena and that this involves being able to predict, control, and answer questions about the phenomena in question. Further, we claimed that such understanding is often model-induced. That is, scientists come to understand phenomena through the construction, use, and analysis of models; this applies to simple equations or box-and-arrow-style models as well as to contemporary deep learning-based models. In this section, we elaborate on how ANNs may contribute to an improved understanding of neuroscientific phenomena. First, we argue that deep learning-based models offer a novel *epistemic perspective* which may render accessible previously inaccessible aspects of neuroscientific phenomena (section 3.1). Second, we argue that understanding complex phenomena typically requires integrating *distributed explanatory information* produced from multiple epistemic perspectives (section 3.2). With this conceptual machinery in place, we sketch a specific research strategy by which ANN-driven neuroscience may contribute to researchers' understanding of biological brains (section 3.3).

3.1 A Novel Epistemic Perspective

Epistemic perspectives are best described as ways in which researchers approach a target phenomenon – through the use of tools, skills, and theoretical assumptions – to yield explanatory information about how it is elicited by a system. Figuratively speaking, we may think of an epistemic perspective as the outlook or view a scientist takes on a phenomenon; it is constrained by their specific knowledge and capacities, as well as their research questions. More specifically, epistemic perspectives can be characterized with respect to the following dimensions: i) spatiotemporal granularity, ii) specificity, i.e., which kinds of entities and dependencies are detected and described, iii) point of view, including ontological and theoretical assumptions, iv) sensitivity to different factors, and v) scope, i.e., class of phenomena which are investigated (Kästner, 2018, p. 74). Crucially, different epistemic perspectives need not be conceived as competing with each other. Although research from a particular perspective is sometimes sufficient for answering specific questions, deeper

understanding of complex systems requires figuring out how insights from multiple epistemic perspectives fit together (Mitchell, 2002, 2019). Before we consider the question of how multiple perspectives can be integrated (section 3.2), let us briefly sketch the core features of the epistemic perspective a deep learning approach to computational neuroscience provides. Needless to say, this will be unavoidably simplistic. In reality, practitioners constructing ANN models of neural functions may actually have different skills and tools and endorse various theoretical commitments. Still, it is valuable to see where the contrast lies to more biologically-grounded neuroscience research.

Spatiotemporal Granularity. Traditional neuroscience methods are constrained with respect to both spatial and temporal resolution, though in various ways (Churchland & Sejnowski, 1988; Grinvald & Hildesheim, 2004). Similarly, the deep learning approach to computational neuroscience can be applied at numerous spatial scales, e.g., by targeting neural populations in specific anatomical regions (Sussillo et al., 2012) or distributed processing along pathways like the ventral visual stream (Yamins et al., 2014). What is novel, though, is that *the same model* can shed light on neural phenomena across a wide range of temporal scales. If training is taken into consideration, one model can be taken to represent processes extending all the way from a single perceptual judgement to evolutionary learning (Cao & Yamins, 2021; Zador, 2019).

Specificity. Being based on automated learning procedures, the deep learning approach has the potential to pick up on patterns in data and dependencies between factors that a human researcher will not usually see or hypothesize. A case in point are functional groups of neurons in V4 we discuss below (section 3.3). For a detailed discussion of how deep learning-based models can help uncover structures inaccessible from other epistemic perspectives see Kästner & Crook (manuscript).

Point of View. The deep learning approach casts the brain as an optimisation machine, with specialised modules solving individual computational problems (Cao & Yamins, 2021; Marblestone et al., 2016). This can be contrasted with more biologically-grounded approaches to neuroscience which reject some of these theoretical assumptions (Buzsáki, 2006; Pessoa, 2023; Pulvermüller et al., 2021).

Sensitivity. Unlike biologically-grounded neuroscience, which may treat any and all features of neuroanatomy as relevant to understanding the brain (including, e.g., glial cells or endocrine signals), the deep learning approach is primarily sensitive to features which are necessary for solving the computational task at hand. Further, because practitioners are interested in how *the brain* solves these tasks, they especially value models which replicate characteristic error profiles in behavioural tasks (say, object recognition) and display similar patterns of neural activity (as measured by, e.g., fMRI) (Bowers et al., 2022; Schrimpf et al., 2020).

Scope. The flexibility of the deep learning approach renders it amenable to shedding light on a broad range of cognitive and perceptual phenomena, including vision (Zhuang et al., 2021), audition (Kell et al., 2018), language comprehension (Schrimpf et al., 2020) working memory (Kozachkov et al., 2022), and many more besides (see Cichy & Kaiser, 2019; Doerig et al., 2023 for discussion). Unlike biologically-grounded neuroscience, this research is not constrained by ethical concerns or the availability of model organisms. Thus, the deep learning approach has a broader scope allowing practitioners to study, for instance, the "evolutionary goals and historical or developmental constraints that are responsible for shaping a system" (Cao & Yamins, 2021, p. 2; see also Lillicrap & Kording, 2019). These dependencies can be illuminated by focusing on the *designed* components of ANNs – objective functions, architecture, and learning algorithms – which do not suffer from problems of opacity (Richards et al., 2019; Thompson, 2021). In addition, we think deep learning can also help uncover the functional and computational structure of (biological as well as artificial) neural networks (section 3.3).

In summary, our exposition makes it plain that the deep learning approach provides a novel epistemic perspective on biological neural processing which renders accessible information unavailable from other epistemic perspectives (i.e., those typically taken by more biologically-grounded neuroscientists). As such, we think it crucially complements existing research practices aiming to understand the brain. To unlock its full potential, however, insights from the deep learning approach must likely be combined with those gained from traditional modelling and biologically-grounded neuroscience.

3.2 Distributed Explanation, Integration, and Constraints

We already hinted that the relationship between understanding models and understanding phenomena is complex. One major reason this is so, we take it, is that *in practice* a subject S's understanding a phenomenon P will not usually rely on just a single model. According to Hochstein, "an individual model is rarely applied in isolation, and is often used to complement a huge body of background information and pre-existing models about the target system" (2016, p. 1401). Recall that for S to understand P involves the abilities to (i) make coarse-grained predictions about P, (ii) give rich, detailed, and productive answers to a broad range of questions concerning P, and (iii) intervene on and control P (see section 2.2). To achieve this all of this, it will usually be required to for S to draw on a *distributed body of information* about P and the system eliciting it.

To clarify, it will be useful to briefly talk about distributed explanations – where the idea of a distributed body of information has been explored. Distributed explanations consist, at least partially, in sets of models, where each model provides different explanatory information about a given phenomenon by illuminating it from different epistemic perspectives (cf., Craver & Kaplan, 2020; Hochstein, 2016; Veit, 2020). If this is correct, philosophers limiting their analyses to "particular models instead of sets of models commit a fatal mistake" (Veit, 2020, p. 108). Individual models only ever yield partial understanding of the phenomenon in question. While we agree that sets of models are key constituents of distributed explanations, we see no reason to stop there. The *explanatory store* which facilitates scientific explanation and – as far as we are concerned – understanding should also admit data graphs, diagrams, explanatory writing in natural language, and any other representational vehicles that support understanding (Burnston, 2016; Kästner, 2018; Kohár & Krickel, 2021).[11] Thus, the items that support scientific understanding (viz. the elements of the explanatory store) may be of very different kinds and, just as models, may come in varying degrees of complexity.

Despite the looming messiness of the overall situation, we think it is important to acknowledge both the distributed character of the information which supports scientific understanding and the varying contributions that different elements in the explanatory store can make to it. First and foremost, this is because it accords with scientific practice. For instance, Craver and Kaplan observe that "many models are involved, explicitly or implicitly,

[11] The term *explanatory store* is originally due to Kitcher (1981) and is also used by Craver and Kaplan (2020).

in most explanations" and that we use various models to "gesture towards explanations that invariably have more content about the causal structure of the world than any single, useful model can express" (2020, pp. 309–310). Besides, once we acknowledge that the information that supports scientific understanding is distributed, we can circumvent undesirable trade-offs inherent to model building (Matthewson & Weisberg, 2009). There is no principled reason we cannot enjoy predictive success *and* improved understanding if we pool insights from multiple models in a common explanatory store.

Unfortunately, these benefits do not come for free. Reaping them requires at least a provisional account of *how* different items in the explanatory store (viz. information generated from multiple epistemic perspectives) can support greater understanding than mastery of particular models (c.f., Chang, 2012, Chapter 5). The answer, we think, is that many items in the explanatory store not only contribute explanatory information but also *constraints* (Lawler & Sullivan, 2021; Sullivan, 2022). Based on these constraints, insights provided by different models (or other items) may be *integrated* with one another (see also Kästner, 2018).

Focusing on models specifically, we might say that each model "provides a partial grasp of the phenomenon, and each requires input and ongoing engagement with the other perspectives" (Mitchell, 2019, p. 179). In our view, the engagement Mitchell describes often consists of *constraining interpretation*. The interpretation of each source of explanatory information is constrained by the rest of the explanatory store (or, for cognitive tractability, a small but relevant subset). The exact way this plays out in any single case is subtle, depending crucially on the evidentiary support different sources of explanatory information enjoy, their perceived relevance to the phenomenon at hand, and what aspects of the phenomenon scientists are particularly interested in. For illustration, we briefly present two cases that highlight ways in which the deep learning approach can support scientific understanding of biological neural processing. Though the cases are different, they each involve contributing insights from a novel epistemic perspective and integrating them with other evidence.

Case 1: Vision. Bowers and colleagues' (2022) make the case that the deep learning-based approach to vision neuroscience has paid inadequate attention to research documenting the idiosyncratic features of biological vision. The failure to adapt computational models to these constraints has resulted in the development of models which fail to replicate observed experimental effects in human vision. For instance, while humans show a clear preference for

shape over texture, vision ANNs show the opposite pattern (Geirhos et al., 2018). As a result, Bowers and colleagues argue, we cannot (yet) utilize these models to gain insights into the algorithmic structure of biological vision. However, in light of evidence showing similar functional specialisation of single neurons in ANNs and biological visual systems (Pospisil et al., 2018; Willeke et al., 2023), we think a more nuanced conclusion should be drawn. Our understanding of human vision should provisionally integrate *some* properties of vision ANNs (e.g., the hierarchical composition of representations and the functional specialisation of individual units), while leaving others out (e.g., early layer neurons' overly strong preference for texture). Then, by systematically varying ANN architectures (e.g., by introducing recurrence, Kar et al., 2019) and training procedures (e.g., switching from supervised to unsupervised learning, Zhuang et al., 2021) and comparing the resulting models with diverse sources of neuroscientific data, researchers can clarify which features of ANNs are really i) biologically plausible, and ii) functionally relevant. This way, neuroscientists can refine predictions about how neural populations will respond to certain stimuli and target loci of control (Bashivan et al., 2019). This effectively means integrating research from the epistemic perspectives of traditional vision neuroscience and the deep learning approach to increase our understanding of biological vision.

Case 2: Working Memory. Kozachov and colleagues (2022) use recurrent neural networks (RNNs) to explore the hypothesis that short-term synaptic plasticity (STSP) plays a functional role in working memory (Mongillo et al., 2008). The authors trained RNNs both with and without an STSP mechanism to maintain items in working memory over a delay period. They then assessed them with respect to three criteria: 1) the robustness of their performance in the face of distractors, 2) graceful degradation to synaptic loss, 3) similarity to the neural dynamics of the prefrontal cortex. Kozachov and colleagues found that both types of RNN were robust to distractors, but that those with the STSP mechanism degraded more gracefully and exhibited dynamics closer to those of the (primate) prefrontal cortex. In other words, the epistemic perspective researchers adopted by applying the deep learning approach allowed them to reveal previously unknown dependencies between architectural and functional properties of neural structures. Put more generally, we might say that the deep learning approach lent credence to particular computational and mechanistic hypotheses about human brain function (e.g., Miller et al., 2018; Mongillo et al., 2008; see also Doerig et al.,

2023). Crucially, in this case, too, it is the integration of insights from different epistemic perspectives that helps researchers advance their understanding of complex phenomena such as working memory.

Thus far, we argued that deep learning-based models offer an additional epistemic perspective for brain researchers that can complement and constrain insights from traditional modelling or biologically-grounded neuroscience by contributing to a distributed store of knowledge. An argument to the same effect could have been made about other innovations in research, e.g. the invention of fMRI scans or the availability of even simple machine learning methods. Against this backdrop readers may wonder: is there anything really *special* about ANN-driven neuroscience?

3.3 Understanding Biological Brains Through Re-engineering

We already argued that we should not fear the bogeyman, but is there reason to welcome him with open arms? We think there is! By combining the strengths of deep learning with post-hoc interpretability techniques and biologically-grounded neuroscience, researchers can access the patterns embedded in trained ANNs and use them to gain insight into the functional and computational structure of neurobiological systems. Notably, the mechanistic description of biological systems this enables is precisely what critics fear might go missing when researchers utilize ML-based models to try and understand the brain (Boge, 2022; Chirimuuta, 2021; López-Rubio & Ratti, 2021; Thompson, 2021). We do not think that the deep learning approach sacrifices functional and computational structures, rather, it sheds a new light on them.

To see why this is and how it might work, recall that deep learning models are usually opaque. While the learned parameters of the models are accessible, understanding how they represent information and implement the mapping from input to output requires a dedicated research effort (i.e., this information is – unlike with traditional scientific models – not *given*, cf. section 2.1). To uncover the functional structure and computational properties of trained ANNs, the application of systematic post-hoc interpretability techniques is needed (e.g., Bau et al., 2017; Cammarata et al., 2020; Geiger et al., 2021, 2022; Nanda et al., 2023; Olah et al., 2018). As of yet (and as far as we are aware), the research area that most vigorously employs post-hoc interpretability techniques is explainable AI (XAI), not neuroscience. While we are

certainly not the only ones to think that XAI presents opportunities for scientific discovery generally (see Zednik & Boelsen, 2022), we believe that applying post-hoc interpretability techniques to trained ANN models provides a unique opportunity for neuroscience by facilitating an *iterative re-engineering process*.

The process we have in mind involves the following key steps. First, train an ANN to implement an ethologically relevant task (as described in Yamins & DiCarlo, 2016). Second, probe the behavioural characteristics of the trained ANN and evaluate the results with respect to explanatory information distributed across the explanatory store (as in Bowers et al., 2022).[12] Third, employ systematic post-hoc interpretability techniques to uncover the functional and computational structure of the trained ANN (as in Cammarata et al., 2020). Fourth, use these findings to develop hypotheses about the target biological system (Willeke et al., 2023). Fifth, use a variety of more traditional neuroscientific methods to test whether those hypotheses apply to biological target systems. Each step in the process may need to be refined and repeated multiple times in light of the constraints provided by the other steps.

The greatest source of uncertainty over the promise of the iterative re-engineering process is the plausibility of the third step, viz. the usefulness of post-hoc interpretability techniques (Cearns et al., 2019). However, recent work by a group around Chris Olah (Cammarata et al., 2020, 2021; Elhage et al., 2021; Olah et al., 2018) serves as a proof of principle that the kind of strategy we have in mind is applicable in practice. The researchers set out to characterise the functional structure of the image classification ANN *InceptionV1* (note that this ANN was not developed explicitly as a model of biological vision). They applied systematic post-hoc interpretability techniques to reverse engineer the opaque system. Focusing on curve detection specifically they found that:

> "although curve detection involves more than 50,000 parameters, those parameters actually implement a simple algorithm that can be read off the weights and described in just a few English sentences." (Cammarata et al., 2021)

Indeed, the detailed understanding of the system's functional structure enabled the researchers to re-design this part of the system (i.e., a curve circuit) from scratch.

[12] Obtaining promising results in this step may be a pre-requisite for engaging in the laborious and challenging process of reverse-engineering an ANN. As such, steps 1 and 2 already form a mini-loop which may be iterated over numerous times before step 3 is reached.

Subsequently, they investigated – with research strategies familiar from biologically-grounded neuroscience – to what extent this re-engineered system actually displayed the same properties as the original system (for a detailed discussion see Kästner & Crook manuscript). This way, they validated their interpretable model of inceptionV1's curve detectors. Though this kind of research is still in its infancy, it clearly illustrates that re-engineering these ANNs through the application of post-hoc interpretability techniques *is possible*. As such, we think doing so on ANNs trained on ethologically relevant tasks might provide a fruitful way to uncover functional and structural features that can be converted into hypotheses for brain research.

Indeed, we know of one case in which this research has already inspired hypothesis generation. After synthesizing images (MEIs) to maximally excite V4 neurons in the macaque visual system, Willeke and colleagues (2023, p. 2) observed that units were arranged into functional groups selective for "specific complex visual features such as eye-like structures, oriented fur patterns, grid-like motifs, or curvatures." Referring to Olah's work, these researchers noted a "striking similarity between [the] single cell MEIs of V4 neurons and single units in the InceptionV1 architecture" (2023, p. 10).[13] Building on this similarity, Willeke et al. state that "the resemblance between V4 neuronal and deep artificial neural network feature selectivity can be used to generate specific hypotheses about visual tuning properties of primate V4 neurons". More specifically, they suggest that ANNs can be used to "derive predictions about color boundary encoding in monkey V4 functional groups, which could subsequently be verified in in vivo experiments" (2023, p. 7). These descriptions capture steps 3-5 of our iterative re-engineering process precisely.

Naturally, which precise models, tools and theories will be utilized most effectively throughout such a process obviously remains an empirical question. Indeed, it will likely vary from case to case. Still, we believe that we have shown that the iterative re-engineering research strategy has great potential. It may improve scientists' ability to make coarse qualitative predictions about brain behaviour, answer questions about biological neural processing, and design

[13] Note that Willeke and colleagues performed psychophysics experiments and quantitative analyses to verify the robustness of their qualitative similarity judgements.

effective interventions on the brain. That is, it provides an opportunity to foster our understanding of the brain in yet unimagined ways.

4 Conclusion

The aim of (neuro)science is to foster understanding of phenomena in the world. Machine learning approaches, particularly deep learning-based ANNs, can contribute to this endeavour by offering a novel and unique epistemic perspective affording insights into the operation of complex systems that are otherwise unavailable. While some have been worrying that this strategy implies trading one black box for another and might actually hinder scientific understanding, we argued these worries are unfounded. There is no trade-off between the predictive success of a model and how much understanding it can confer. Thus, utilizing complex computational models in neuroscience will not generally inhibit our ability to understand the (human) brain. Quite to the contrary, we believe, deep learning is best conceived as complementary to established neuroscientific methodology. It offers a powerful addition to neuroscientists' toolkit enabling constraint-based integration of numerous complementary epistemic perspectives.

We thus conclude that deep learning contributes to – rather than detracts from – understanding the phenomena neuroscientists investigate. While our argument has focused on the domain of neuroscience specifically, we suggest that it is likely to carry over to relevantly similar scientific domains, even if the details may differ.

Acknowledgements

Work on this paper has been supported by the project "Explainable Intelligent Systems (EIS)" founded by the Volkswagen Foundation (Az 9B830). The authors wish to thank all project members for their discussion of an early version of this paper. We are especially indebted to Astrid Schomäcker, Sara Mann and Andreas-Sesing-Wagenpfeil for ample constructive feedback.

References

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science (New York, N.Y.)*, *364*(6439), eaav9436. https://doi.org/10.1126/science.aav9436

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). *Network Dissection: Quantifying Interpretability of Deep Visual Representations*. 3319–3327. https://doi.org/10.1109/CVPR.2017.354

Bechtel, W., & Richardson, R. (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. https://doi.org/10.7551/mitpress/8328.001.0001

Bennett, D., Silverstein, S. M., & Niv, Y. (2019). The Two Cultures of Computational Psychiatry. *JAMA Psychiatry*, *76*(6), 563–564. https://doi.org/10.1001/jamapsychiatry.2019.0231

Boge, F. J. (2022). Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines*, *32*(1), 43–75. https://doi.org/10.1007/s11023-021-09569-4

Bokulich, A. (2017). Models and Explanation. In L. Magnani & T. Bertolotti (Eds.), *Springer Handbook of Model-Based Science* (pp. 103–118). Springer International Publishing. https://doi.org/10.1007/978-3-319-30526-4_4

Boon, M. (2020). How Scientists Are Brought Back into Science—The Error of Empiricism. *A Critical Reflection on Automated Science: Will Science Remain Human?*, 43–65. https://doi.org/10.1007/978-3-030-25001-0_4

Bouatta, N., Sorger, P., & AlQuraishi, M. (2021). Protein structure prediction by AlphaFold2: Are attention and symmetries all you need? *Acta Crystallographica. Section D, Structural Biology*, *77*(Pt 8), 982–991. https://doi.org/10.1107/S2059798321007531

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2022). Deep Problems with Neural Network Models of Human Vision. *The Behavioral and Brain Sciences*, 1–74. https://doi.org/10.1017/S0140525X22002813

Burnston, D. C. (2016). Data graphs and mechanistic explanation. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *57*, 1–12. https://doi.org/10.1016/j.shpsc.2016.01.002

Buzsáki, G. (2006). *Rhythms of the Brain*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195301069.001.0001

Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., & Schubert, L. (2020). Thread: Circuits. *Distill*, *5*(3), 10.23915/distill.00024. https://doi.org/10.23915/distill.00024

Cammarata, N., Goh, G., Carter, S., Voss, C., Schubert, L., & Olah, C. (2021). Curve Circuits. *Distill*, *6*(1), e00024.006. https://doi.org/10.23915/distill.00024.006

Cao, R., & Yamins, D. (2021). Explanatory models in neuroscience: Part 2 -- constraint-based intelligibility. *ArXiv:2104.01489 [Cs, q-Bio]*. http://arxiv.org/abs/2104.01489

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, *9*(1), Article 1. https://doi.org/10.1038/s41398-019-0607-2

Chang, H. (2012). *Is Water H2O? Evidence, Realism and Pluralism*. Boston Studies in the Philosophy and History of Science.

Charles Leek, E., Leonardis, A., & Heinke, D. (2022). Deep neural networks and image classification in biological vision. *Vision Research*, *197*, 108058. https://doi.org/10.1016/j.visres.2022.108058

Chirimuuta, M. (2021). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*, *199*(1), 767–790. https://doi.org/10.1007/s11229-020-02713-0

Chollet, F. (2021). *Deep Learning with Python, Second Edition*. Simon and Schuster.

Churchland, P. S., & Sejnowski, T. J. (1988). Perspectives on cognitive neuroscience. *Science (New York, N.Y.)*, *242*(4879), 741–745. https://doi.org/10.1126/science.3055294

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Colaco, D. (2019). *An Investigation of Scientific Phenomena* [PhD Thesis]. University of Pittsburgh.

Craver, C. F. (2001). Role Functions, Mechanisms, and Hierarchy. *Philosophy of Science*, *68*(1), 53–74.

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.

Craver, C. F., & Darden, L. (2013). *In Search of Mechanisms: Discoveries across the Life Sciences*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/I/bo16123713.html

Craver, C. F., & Kaplan, D. M. (2020). Are More Details Better? On the Norms of Completeness for Mechanistic Explanations. *The British Journal for the Philosophy of Science*, *71*(1), 287–319. https://doi.org/10.1093/bjps/axy015

Creel, K. A. (2020). Transparency in Complex Computational Systems. *Philosophy of Science*, *87*(4), 568–589. https://doi.org/10.1086/709729

de Regt, H. (2009). The Epistemic Value of Understanding. *Philosophy of Science*, *76*(5), 585–597. https://doi.org/10.1086/605795

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 1–20. https://doi.org/10.1038/s41583-023-00705-w

Elgin, C. (2007). Understanding and the facts. *Philosophical Studies*, *132*(1), 33–42. https://doi.org/10.1007/s11098-006-9054-z

Elgin, C. (2017). *True enough*. MIT Press.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., … Olah, C. (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.

Feest, U. (2017). Phenomena and Objects of Research in the Cognitive and Behavioral Sciences. *Philosophy of Science*, *84*(5), 1165–1176.

Frigg, R., & Hartmann, S. (2020). Models in Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/models-science/

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal Abstractions of Neural Networks. *Advances in Neural Information Processing Systems*, *34*, 9574–9586. https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html

Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., Goodman, N., & Potts, C. (2022). Inducing Causal Structure for Interpretable Neural Networks. *Proceedings of the 39th International Conference on Machine Learning*, 7324–7338. https://proceedings.mlr.press/v162/geiger22a.html

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018, December 21). *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness*. International Conference on Learning Representations. https://openreview.net/forum?id=Bygh9j09KX

Glaser, J. I., Benjamin, A. S., Farhoodi, R., & Kording, K. P. (2019). The roles of supervised machine learning in systems neuroscience. *Progress in Neurobiology*, *175*, 126–137. https://doi.org/10.1016/j.pneurobio.2019.01.008

Glennan, S. (2017). *The new mechanical philosophy* (First edition). Oxford University Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Grimm, S. (2021). Understanding. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/understanding/

Grinvald, A., & Hildesheim, R. (2004). VSDI: A new era in functional imaging of cortical dynamics. *Nature Reviews. Neuroscience*, *5*(11), 874–885. https://doi.org/10.1038/nrn1536

Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, *40*(2), Article 2. https://doi.org/10.1609/aimag.v40i2.2850

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Hills, A. (2016). Understanding Why. *Noûs*, *50*(4), 661–688. https://doi.org/10.1111/nous.12092

Hochstein, E. (2016). One mechanism, many models: A distributed theory of mechanistic explanation. *Synthese*, *193*(5), 1387–1407. https://doi.org/10.1007/s11229-015-0844-8

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, *169*(3), 615–626. https://doi.org/10.1007/s11229-008-9435-2

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), Article 7873. https://doi.org/10.1038/s41586-021-03819-2

Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, *46*(3), 240–254. https://doi.org/10.1016/j.tins.2022.12.008

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), Article 6. https://doi.org/10.1038/s41593-019-0392-5

Kästner, L. (2018). Integrating mechanistic explanations through epistemic perspectives. *Studies in History and Philosophy of Science Part A*, *68*, 68–79. https://doi.org/10.1016/j.shpsa.2018.01.011

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses,

and Reveals a Cortical Processing Hierarchy. *Neuron*, *98*(3), 630-644.e16. https://doi.org/10.1016/j.neuron.2018.03.044

Kitcher, P. (1981). Explanatory Unification. *Philosophy of Science*, *48*(4), 507–531.

Kohár, M., & Krickel, B. (2021). Compare and Contrast: How to Assess the Completeness of Mechanistic Explanation. In F. Calzavarini & M. Viola (Eds.), *Neural Mechanisms: New Challenges in the Philosophy of Neuroscience* (pp. 395–424). Springer International Publishing. https://doi.org/10.1007/978-3-030-54092-0_17

Kozachkov, L., Tauber, J., Lundqvist, M., Brincat, S. L., Slotine, J.-J., & Miller, E. K. (2022). Robust and brain-like working memory through short-term synaptic plasticity. *PLOS Computational Biology*, *18*(12), e1010776. https://doi.org/10.1371/journal.pcbi.1010776

Lawler, I., & Sullivan, E. (2021). Model Explanation Versus Model-Induced Explanation. *Foundations of Science*, *26*(4), 1049–1074. https://doi.org/10.1007/s10699-020-09649-1

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), Article 7553. https://doi.org/10.1038/nature14539

Lillicrap, T. P., & Kording, K. P. (2019). *What does it mean to understand a neural network?* (arXiv:1907.06374). arXiv. https://doi.org/10.48550/arXiv.1907.06374

Lindsay, G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031. https://doi.org/10.1162/jocn_a_01544

López-Rubio, E., & Ratti, E. (2021). Data science and molecular biology: Prediction and mechanistic explanation. *Synthese*, *198*(4), 3131–3156. https://doi.org/10.1007/s11229-019-02271-0

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an Integration of Deep Learning and Neuroscience. *Frontiers in Computational Neuroscience*, *10*. https://www.frontiersin.org/article/10.3389/fncom.2016.00094

Matthewson, J., & Weisberg, M. (2009). The structure of tradeoffs in model building. *Synthese*, *170*(1), 169–190. https://doi.org/10.1007/s11229-008-9366-y

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working Memory 2.0. *Neuron*, *100*(2), 463–475. https://doi.org/10.1016/j.neuron.2018.09.023

Mitchell, S. D. (2002). Integrative Pluralism. *Biology and Philosophy*, *17*(1), 55–70. https://doi.org/10.1023/A:1012990030867

Mitchell, S. D. (2019). Perspectives, Representation, and Integration. In *Understanding Perspectivism* (pp. 178–193). Routledge. https://doi.org/10.4324/9781315145198-11

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic Theory of Working Memory. *Science*, *319*(5869), 1543–1546. https://doi.org/10.1126/science.1150769

Morgan, M. S., & Morrison, M. (1999). *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge University Press.

Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). *Progress measures for grokking via mechanistic interpretability* (arXiv:2301.05217). arXiv. https://doi.org/10.48550/arXiv.2301.05217

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The Building Blocks of Interpretability. *Distill*, *3*(3), 10.23915/distill.00010. https://doi.org/10.23915/distill.00010

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2. ed). Cambridge Univ. Press.

Pessoa, L. (2023). The Entangled Brain. *Journal of Cognitive Neuroscience*, *35*(3), 349–360. https://doi.org/10.1162/jocn_a_01908

Pospisil, D. A., Pasupathy, A., & Bair, W. (2018). 'Artiphysiology' reveals V4-like shape tuning in a deep network trained for image classification. *ELife*, *7*, e38242. https://doi.org/10.7554/eLife.38242

Potochnik, A. (2016). *Scientific Explanation: Putting Communication First*. 721–732.

Pulvermüller, F., Tomasello, R., Henningsen-Schomers, M. R., & Wennekers, T. (2021). Biological constraints on neural network models of cognitive function. *Nature Reviews Neuroscience*, *22*(8), Article 8. https://doi.org/10.1038/s41583-021-00473-5

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. https://doi.org/10.1038/s41593-019-0520-2

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, *108*(3), 413–423. https://doi.org/10.1016/j.neuron.2020.07.040

Shagrir, O., & Bechtel, W. (2017). Marr's Computational Level and Delineating Phenomena. In D. M. Kaplan (Ed.), *Explanation and Integration in Mind and Brain Science* (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780199685509.003.0009

Srećković, S., Berber, A., & Filipović, N. (2022). The Automated Laplacean Demon: How ML Challenges Our Views on Prediction and Explanation. *Minds and Machines*, *32*(1), 159–183. https://doi.org/10.1007/s11023-021-09575-6

Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A*, *44*(3), 510–515. https://doi.org/10.1016/j.shpsa.2012.12.005

Sullivan, E. (2022). Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*, *73*(1), 109–133. https://doi.org/10.1093/bjps/axz035

Sussillo, D., Nuyujukian, P., Fan, J. M., Kao, J. C., Stavisky, S. D., Ryu, S., & Shenoy, K. (2012). A recurrent neural network for closed-loop intracortical brain-machine interface decoders. *Journal of Neural Engineering*, *9*(2), 026027. https://doi.org/10.1088/1741-2560/9/2/026027

Thompson, J. A. F. (2021). Forms of explanation and understanding for neuroscience and artificial intelligence. *Journal of Neurophysiology*, *126*(6), 1860–1874. https://doi.org/10.1152/jn.00195.2021

Veit, W. (2020). Model Pluralism. *Philosophy of the Social Sciences*, *50*(2), 91–114. https://doi.org/10.1177/0048393119894897

Willeke, K. F., Restivo, K., Franke, K., Nix, A. F., Cadena, S. A., Shinn, T., Nealley, C., Rodriguez, G., Patel, S., Ecker, A. S., Sinz, F. H., & Tolias, A. S. (2023). *Deep learning-driven characterization of single cell tuning in primate visual area V4 unveils topological organization* (p. 2023.05.12.540591). bioRxiv. https://doi.org/10.1101/2023.05.12.540591

Woodward, J. F. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, *10*(1), 3770. https://doi.org/10.1038/s41467-019-11786-6

Zednik, C., & Boelsen, H. (2022). Scientific Exploration and Explainable Artificial Intelligence. *Minds and Machines*, *32*(1), 219–239. https://doi.org/10.1007/s11023-021-09583-6

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118. https://doi.org/10.1073/pnas.2014196118