J Neurophysiol 126: 1860–1874, 2021.
First published October 13, 2021; doi:10.1152/jn.00195.2021

american physiological society

JNP JOURNAL OF NEUROPHYSIOLOGY

THE NEUROCENE

Computational Neuroscience

Forms of explanation and understanding for neuroscience and artificial intelligence

Jessica A. F. Thompson

Human Information Processing Lab, Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom

Abstract

Much of the controversy evoked by the use of deep neural networks as models of biological neural systems amount to debates over what constitutes scientific progress in neuroscience. To discuss what constitutes scientific progress, one must have a goal in mind (progress toward what?). One such long-term goal is to produce scientific explanations of intelligent capacities (e.g., object recognition, relational reasoning). I argue that the most pressing philosophical questions at the intersection of neuroscience and artificial intelligence are ultimately concerned with defining the phenomena to be explained and with what constitute valid explanations of such phenomena. I propose that a foundation in the philosophy of scientific explanation and understanding can scaffold future discussions about how an integrated science of intelligence might progress. Toward this vision, I review relevant theories of scientific explanation and discuss strategies for unifying the scientific goals of neuroscience and AI.

causality; deep learning; explanation; intelligibility; understanding

INTRODUCTION

I completed my graduate studies in cognitive computational neuroscience[1] during the deep learning revolution in machine learning and what is now being called the artificial intelligence (AI) revolution in neuroscience (1). My first time attending the Neural Information Processing Systems Conference was in 2011, 1 yr before Geoff Hinton presented the AlexNet paper (2), which marked the end of the most recent AI winter and the beginning of the current deep learning era of machine learning. As deep learning approaches continued to demonstrate their ability to perform a variety of computer vision, machine hearing, and language processing tasks, several neuroscientists came to the same conclusion: that these machine learning systems had potential to be useful as models of biological neural systems[2] (3).

Early empirical work inspired by this potential characterized a correspondence between convolutional neural networks trained to recognize objects in images and brain regions along the ventral stream of the primate visual system (4–9). Further work in a similar vein has continued to characterize how various aspects of model design or hyperparameters (network architecture, learning rules, cost functions, tasks, datasets, etc.) affect this correspondence with neural and behavioral measurements (10–19) and to what extent similar correspondences can be found in other species and sensory systems (20–24). These works are part of a much broader alignment developing between the fields of neuroscience and AI which includes work on cognitively inspired AI, biologically plausible learning algorithms, neural dynamics, representational geometry, and decision making (25–33).

However, not everyone sees the promise of this alignment between neuroscience and AI, and even among those who do, there is disagreement on the value of specific results and approaches (34–54). This lack of clarity felt especially acute during my doctoral research. I was extremely lucky to be able to collaborate with machine learning researchers, engineers, and neuroscientists in both academia and industry to study information processing in brains and machines. Many

[1]Cognitive computational neuroscience refers to the intersection of cognitive science, artificial intelligence (AI), and neuroscience concerned with the goal to "identify the computational principles that underlie perception, action, and cognition" (1).

[2]There have been many interactions between neuroscience and AI research in the past (see Ref. 3), but the deep learning revolution undoubtedly brought about a new era in this intertwined history.

Correspondence: J. A. F. Thompson (jessica.thompson@psy.ox.ac.uk).
Submitted 27 April 2021 / Revised 2 September 2021 / Accepted 6 October 2021

junior scientists go through a period of questioning or a crisis of purpose at some point during their training. Perhaps especially because of the diverse background assumptions of my collaborators, I was encouraged to question the most basic aspects of my research. Why was I doing this research? To what scientific goal would my research contribute? How were the proposed methods suited to that goal? I found it very challenging to situate my research into a broader scientific enterprise. I could not easily describe how what I was working on would somehow bring my community closer to some ultimate goal. I wanted to "discover the computational mechanisms underlying auditory perception" but I couldn't precisely tell you what that meant or how I would know whether I had succeeded at taking a useful step toward that goal.

My response to this crisis of purpose was to consume all the perspective pieces and discussion panels I could find that addressed the developing tension about machine learning models of biological neural systems and their cognitive capacities. This eventually led me to read more philosophy of science, especially about explanation and understanding. Learning about these topics, which are not typically covered in science training programs, made it easier to step back to see the bigger picture. Scientists, compared with philosophers, may generally be more concerned with the finer details of science: the outcomes of specific experiments, an individual's research program, or the verification of a particular hypothesized causal relationship. This is unsurprising, especially given the current incentive structures in science which reward individuals on short time scales rather than communities on long time scales. But science is ultimately a slow, social project, not an individual one. History and philosophy of science, science studies, and metascience can provide insight into how the fruits of science are produced by scientific activities over long time scales.

This learning journey made me realize that my field is at risk of reinventing the wheel when it comes to these discussions about the merits of different scientific approaches and models. The questions at the center of these discussions (What ought to be the epistemic goals of neuroscience? What does it mean to understand a cognitive or neural phenomenon? What is the structure of successful scientific explanations? What does it mean for a model to be intelligible?), are ultimately philosophical in nature. Proper treatment of these important questions will benefit from engagement with the extensive philosophical literature on scientific explanation and understanding. Many neuroscientists may be compelled by these fundamental questions about how to explain neural and cognitive phenomena and be unaware that there exists a whole subfield of philosophy of science dedicated to the philosophy of neuroscience, in which detailed models have been proposed to account for scientific explanation in neuroscience. Let us build on what has already been done rather than starting anew.

You may be thinking at this point, "oh great, not only do I need to learn about both neuroscience and machine learning to follow contemporary computational neuroscience, now you're telling me I need to know about philosophy of science too?!" I know, I know. It is terrible. We cannot be expected to know everything, can we?! Which is partly why I was motivated to write this essay to highlight some of the philosophy of science that I have found most useful for contextualizing recent developments at the intersection of neuroscience and AI and to make some modest proposals for how we might formalize our explanatory goals. I write as a scientist who builds machine learning models, conducts human experiments, and wants to understand cognition and perception. I am not an expert in philosophy of science, and thus I cannot write an authoritative review. Such reviews exist (see Refs. 55–57), but they are largely not written for scientists, nor do they fit in a single essay. So, consider this an opinionated, idiosyncratic highlights reel, written for my peers in cognitive computational neuroscience.

In what follows, I will try to show how philosophy of science, and in particular, theories of explanation, can illuminate some of the core issues currently facing neuroscience and AI. In the next section, I argue that contemporary debates about the merits of AI models in neuroscience amount to debates over what constitutes scientific progress. As such, one must first define the goal(s) toward which one wants to make progress. This motivates the focus on the goals of understanding and explanation developed in EPISTEMIC GOALS: UNDERSTANDING AND EXPLANATION. THEORIES OF SCIENTIFIC EXPLANATION is the review part of the paper, where I introduce a handful of theories of scientific explanation that seem especially relevant to cognitive and computational neuroscience. In SCIENTIFIC EXPLANATION AT THE INTERSECTION OF NEUROSCIENCE AND AI, I further develop my own positions on how to approach scientific explanation at the intersection of neuroscience and AI. Last, in REAPING THE BENEFITS OF AN INTERDISCIPLINARY COGNITIVE COMPUTATIONAL NEUROSCIENCE, I will discuss how the interdisciplinarity of cognitive computational neuroscience, and therefore its diversity of background assumptions, has the potential to facilitate increased objectivity. I hope that this work can scaffold future discussions about how an integrated science of intelligence might progress.

DEBATES ABOUT THE MERITS OF AI MODELS IN NEUROSCIENCE AMOUNT TO DEBATES OVER WHAT CONSTITUTES SCIENTIFIC PROGRESS

Proponents of AI models in neuroscience must often respond to the claim that modeling neural systems with deep neural networks amounts to "replacing one black box with another." This phrase communicates the concern that even though a model might help to capture more variance in some neural activity, that it does not necessarily bring us any closer to our ultimate scientific goals. Thus, we must first ask, progress toward what? To what scientific goals do AI models in neuroscience contribute and how do they facilitate progress?

Traditionally, research on computational modeling of sensory systems has built models that embody specific hypotheses about the nature of information processing at work in some part of the brain. For example, in studying the auditory system, we might be interested in whether (or where in the brain) temporal and spectral information are coded independently or jointly. We could build several models of the

response properties of individual neurons (or voxels) that reflect our hypotheses (temporal, spectral, spectro-temporal). By comparing these models, either via encoding analysis or representational similarity analysis, we can identify neurons (or voxels or regions) where the responses are better captured when conditioning on both spectral and temporal features than when only conditioning on either alone (58).

However, say we trained an artificial neural network to classify natural sounds and then used the resulting learned representations as features in our encoding analysis. We might find these learned representations better predict the collected neural activity than even our best spectro-temporal model (as was found in Ref. 20). As the intermediate features of the neural network were learned rather than designed, we do not know a priori what they "represent." Thus, the critic says, without further characterization of the network itself, this comparison cannot provide insight into what features the auditory system is using when it processes sounds, i.e., the signal processing that it performs. Instead of one thing we do not understand (information processing in the auditory brain), now we have two things we do not understand (information processing in the auditory brain and information processing in a neural network trained to classify sounds). According to the critic, we might have a highly predictive model, but the original scientific questions are left unanswered[3] (44).

Further analysis of this situation and of the value of the neural network model in the example just described must necessarily refer to scientific goals and the means by which one might accomplish those goals. Proponents might seek to describe scientific goals to which AI models can contribute. For example, one might define the goal of understanding as "being able to build." On the other hand, the critic will need to justify why predictive accuracy, variance explained or representational similarity, which were acceptable proxies for the value of hand-designed process models, do not indicate the value of neural network models. For example, they might claim that neural network models are "uninterpretable" or remind us that the goal of modeling in science is not merely faithful recapitulation, otherwise "the best material model for a cat would be another, or preferably the same, cat" (59, p. 320). The notion of black box will be relative to a particular epistemic goal or query. Neural networks are completely transparent and scrutable in one sense; all of their parameters and activities are accessible. We know the precise mathematical function that any given network is implementing. That we feel like there is still more to be known about neural network models despite these facts points to additional epistemic goals for which they remain to some extent opaque.

EPISTEMIC GOALS: UNDERSTANDING AND EXPLANATION

There are a variety of possible goals in science: understanding, explanation, truth, prediction, control, remediation. Here I want to focus on understanding and explanation, two concepts which appear often in contemporary debates in neuroscience and for which there exists a large philosophical literature.

In philosophy, understanding has been discussed as a type of cognitive achievement state (60). For example, when an individual comes to understand a language, another person, a proof, or a scientific theory, that individual has transitioned from a cognitive state of not understanding to understanding. One may distinguish here between understanding, a cognitive achievement state, and the sense of understanding, the subjective experience of understanding. One may also speak of understanding as an accomplishment of a group, where a scientific field, for example, might be said to understand a phenomenon or object of study.

Scientific understanding is distinct from scientific explanation, although exactly how distinct depends on the philosopher. An explanation is typically thought to be a special type of description of a phenomenon to be explained. Explanations can be good or bad (sometimes equated with true or false[4]) and their goodness is primarily a function of some objective relationship to the phenomena they are supposed to explain. Philosophers analyze and theorize about what makes a good scientific explanation. Such theories will typically define the criteria that must be met for a description to be explanatory, and in doing so, distinguish explanations from nonexplanations. Sometimes this second contrast is presented as the difference between explanation and mere description. For example, a set of claims about the appearance of a particular animal species may be accurate and supported by evidence without being explanatory in any way. They are "merely" descriptive.

Scientific understanding has sometimes been used almost interchangeably with scientific explanation by defining scientific understanding as the cognitive state yielded by the possession of a satisfactory explanation. According to this view, the sense of understanding is completely subjective, and hence irrelevant to serious philosophy of science, whereas genuine scientific understanding is something objective which comes about in virtue of a true scientific explanation (61). For example, Craver writes, "All scientists are motivated in part by the pleasure of understanding. Unfortunately, the pleasure of understanding is often indistinguishable from the pleasure of misunderstanding. The sense of understanding is at best an unreliable indicator of the quality and depth of an explanation" (62, p. 21).

Explanations are said to answer why-questions. These why-questions will typically be of the form "why does P?," where P is the phenomenon to be explained, e.g., why does the pendulum swing? An explanation refers to a particular phenomenon, not an object or system. So when we talk about "explaining the brain," we might interpret this as shorthand for "explain all the phenomena the brain is involved in?" Explanation is not synonymous with theory. A scientific theory of a system may help to explain several phenomena.

Classical theories of scientific explanation typically do not include any reference to psychological, social, or pragmatic factors, preferring to focus instead on some objective

[3]This is similar to the criticism outlined in Ref. 44.

[4]I will use *truth* throughout the rest of the text for simplicity, not intending to make any commitments with respect to realism. The reader is free to replace *truth* with *utility* or *epistemic adequacy* if preferred.

relationship between a phenomenon and its explanation. Salmon (63) writes,

> First, we must surely require that there be some sort of *objective* relationship between the explanatory facts and the fact-tobe-explained. Second, not only is there the danger that people will feel satisfied with scientifically defective explanations; there is also the risk that they will be unsatisfied with legitimate scientific explanations. The psychological interpretation of scientific explanation is patently inadequate. (63, p. 13)

Other philosophers have argued for the need for a theory of scientific understanding in its own right—one that is independent of the particular explanatory strategies employed and which does not omit the social, psychological, and pragmatic factors that influence scientific understanding. De Regt (64) distinguishes between phenomenology of understanding (PU), understanding a theory (UT), and understanding a phenomenon (UP). PU and UP map essentially on to concepts already introduced; PU is the subjective sense of understanding and UP is the cognitive state that results from the possession of an adequate explanation of a phenomenon. UT is the ability to use a theory, which De Regt largely associates with intelligibility, "the value that scientists attribute to the cluster of qualities that facilitate the use of the theory" (64, p. 12). De Regt argues that UT is necessary for UP and that there are several pragmatic factors that will influence UT (and therefore UP). For example, certain skills may be required for a scientist to use a theory. Standards of intelligibility are context-dependent and change over time. A theory or model that is deemed unintelligible may become intelligible with time and other scientific developments. This distinction between explanation and understanding is important for analyzing instances where science made mistakes and embraced incorrect explanations. Although our understanding might evolve over time and be a function of our skills and preconceptions, the veracity of an explanation is unchanging (unless the phenomenon changes, of course). UT may be necessary for UP, but it is certainly not sufficient. One may fully understand and be able to use an incorrect theory.

Scientific Understanding Is More than Merely the Subjective Sense of Having Understood

I have noticed that much of the conversation about scientific progress in cognitive and computational neuroscience has focused on understanding. In these conversations, it is crucial that we do not confuse the various notions of understanding and explanation. In particular, it is important that we do not mistake the sense of understanding, the subjective experience of having understood something, for genuine explanatory understanding.

Lillicrap and Kording (65) use "understanding" similarly to how it has been described here, associating it with a cognitive achievement. They emphasize compactness and compressibility in their proposal for what it means to understand a neural network because humans are only able to argue about compact systems. According to their view, any meaningful understanding of a neural network must be compressible into an amount of information that a human can consume, e.g., a textbook. The limits of human cognitive abilities constrain the understanding that can reasonably be sought. For example, humans cannot conceptualize the interactions of 100 trillion synapses simultaneously and so our scientific goals should not require such a feat. To map to De Regt's terminology, Lillicrap and Kording (65) seem to be concerned with UT, not UP nor explanation.

In my view, precisely because, as De Regt emphasizes, UT is context-dependent and pragmatic, we must be somewhat loose with our requirement for UT. Just because a theory seems unintelligible to a particular group of scientists at a specific time and place, we should not infer that it will always be so. For example, deep neural networks as models of cognitive and neural phenomena are becoming more intelligible as we clarify the various research questions to be asked and the space of possible answers and as we develop new empirical, theoretical, and conceptual tools to study AI behavior. Research in deep learning theory and understanding deep learning makes deep networks more intelligible, increasing their potential for use as models of cognition and neural computation.

A bias toward the more intelligible will not necessarily bring us closer to truth and may in fact steer us away from innovative strategies. For example, simplicity is often cited as a marker of intelligibility and a virtue of scientific theories. Historically, science's greatest successes have been the result of describing complex systems with relatively few parameters. However, some neuroscientists have questioned whether this will ever be the case for understanding some neural phenomena (40, 65). The true explanation might be complex, "how could a fixed bias toward simplicity indicate the possibly complex truth any better than a broken thermometer that always reads zero can indicate the temperature? You don't have to be a card-carrying skeptic to wonder what the tacit connection between simplicity and truth-finding could possibly be" (66). I maintain that UP, not UT, is the primary epistemic goal of science and that we may see UT as a weighted constraint (or regularizer[5]) on our search for UP.[6] For an explanation to have impact, it must be intelligible to at least some people, eventually. However, the set of true explanations is not limited by notions of human intelligibility. If an alien race sent Earth a message containing a true explanation, it would still be true even if no human gleaned understanding from it.

[5]A regularizer is a term added to an objective function to bias an optimization toward a particular class of solutions, e.g., to encourage sparsity or smoothness.

[6]This is in line with Woodward's perspective on understanding as a constraint on what an explanation is (67) and is also how I think about other scientific virtues like reproducibility. If we optimize directly for reproducibility, we may end up answering very uninteresting questions whose answers are highly reproducible. The distinction between primary goals and secondary constraints is well captured by analogy to optimization. In the case of understanding and intelligibility, De Regt tells us that the loss landscape is nonstationary.

Certainly we do want intelligible theories and understanding, but we do not only want understanding—we want our understanding to be robust. Thus, if we want to effectively debate how our science will progress, we must consider the explanations we will produce not just the understanding they will provide.

THEORIES OF SCIENTIFIC EXPLANATION

I will now review several theories of scientific explanation, focusing on three broad classes that seem especially relevant to cognitive computational neuroscience: causal explanation, functional explanation, and minimal model explanation[7] (61, 68, 69, 70).

Causal Theories of Explanation

Causal theories of explanation reflect the idea that to explain a phenomenon is to describe its causes, or in other words, of situating a phenomenon in the causal structure of the world.

A popular subtype of causal explanation, especially among philosophers of neuroscience, is mechanistic explanation. According to this view, an explanation will be a description of the mechanisms—the physical entities, activities, and their organization—that produce or realize the phenomenon-to-be-explained (62, 63, 71). A canonical example of causal mechanical explanation in neuroscience is the explanation of the action potential. The textbook explanation of an action potential describes the mechanism by which it occurs. It will describe the component parts (ion channels, synapses, membranes, ions, etc.) and how the action potential is realized by their activities and organization (e.g., ion channels in the cell membrane opening and closing to modulate the membrane potential). The causal interplay between the mechanism components, their activities and organization, explains the occurrence of the action potential.

When applied to computational neuroscience, this view has led to the model-to-mechanism mapping (3M) model. According to 3M, a computational model is explanatory only when it satisfies the following requirements:

1) the variables in the model correspond to identifiable components and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon,
2) the causal relations posited among these variables in the model correspond to the activities or operations among the components of the target mechanism (72, p. 272).

In 3M, a computational model is explanatory to the extent that it faithfully describes the mechanisms that realize the phenomenon to be explained (71).

Causal mechanists sometimes distinguish between etiological and constitutive explanations (and mechanisms), which differ in whether the phenomenon is explained by tracing the causal story that led to its occurrence or via a causal analysis of the phenomenon itself (73). I introduce this distinction here because I will use it later to situate the deep learning framework for neuroscience.

1) Etiological explanation: to explain in terms of antecedent causes, i.e., to "trace the causal processes and interactions leading up to [the event]" (74, p. 44), e.g., the virus causes the flu, dehydration causes thirst. The target phenomenon is produced by the mechanism.
2) Constitutive (or componential) explanation: to explain via description of causal relationships among component parts and their activities. The target phenomenon is realized by the mechanism.

Salmon suggests that "[casual explanations can] possess both etiological and constitutive aspects. To explain the destruction of Hiroshima by a nuclear bomb, we need to explain the nature of a chain reaction (constitutive aspect) and how the bomb was transported by airplane, dropped, and detonated (etiological aspect)" (73, p. 324). We can further distinguish between structural and triggering etiological mechanisms. Structural mechanisms set up the necessary conditions such that a trigger will cause the target phenomenon to occur. Structural mechanisms can be selective (like natural selection) or instructive (like pedagogy) (75). Fn7

However, not all casual theories of explanation focus on mechanisms. For example, Woodward's interventionist account says that to explain a phenomenon is to show what that phenomenon depends on. Explanations are patterns of counterfactual dependence describing how the system whose behavior we wish to explain would change under various conditions. These patterns help to answer what-if-things-had-been-different questions. Explanations tell us how we might intervene on the system, what manipulation could be carried out to bring about the antecedent of a counterfactual question (67, 76).

Functional Explanation

A possible alternative to causal explanation is functional explanation. According to Cummins (77), the main phenomena to be explained in psychology are *capacities*: "the capacity to see depth, to learn and speak a language, to plan, to predict the future, to empathize, to fathom the mental states of others, to deceive oneself, to be self-aware, and so on" (78, p. 8–9). He proposes that capacities are explained via functional analysis and realization. Functional analysis refers to the process of decomposing a capacity into progressively simpler subcapacities and their functional organization. Realization in this context refers to the requirement that the analysis must show how the behavior of the parts of the system come together to enable the system to demonstrate the capacity to be explained (79). Cummins uses the example of assembly line production to illustrate:

> Production is broken down into a number of distinct and relatively simple (unskilled) tasks. The line has the capacity to produce the product in virtue of the fact that the units on the line have the capacity to perform one or more of these tasks, and in virtue of the fact that when these tasks are performed in a certain organized way—according to a certain program—the finished product results. (78, p. 12)

[7]For interested readers, other approaches to explanation that I omitted for brevity include deductive nomological (61), statistical relevance (68), unificationist (69), and pragmatic (70).

Realization could take the form of building a computational model according to functional constraints and showing that it exhibits the capacity of interest. On this view, the role of neuroscience in the explanation of psychological capacities might be to arbitrate among competing functional analyses, which are themselves formulated in nonneural terms.

Functional explanation has received considerable criticism from causal mechanists, who claim that functional analyses provide incomplete characterizations of mechanisms, or *mechanism sketches* (80). These sketches are only explanatory to the extent that they satisfy the constraints of causal mechanical explanation. In this sense, functional analyses, such as those performed in cognitive psychology to decompose a cognitive capacity into subcapacities, constitute a first step toward ultimate mechanical explanations (72, 80). As such, a functional description provides only a *how-possibly* model; work remains to be done to turn it into a how-actually model (80). In this light, the localization goal of much of cognitive neuroscience can be viewed as mapping the component operations, as provided by cognitive psychology via functional analysis, on to the component parts, the brain regions (81).

In contrast, the functionalist perspective states that functional explanations in psychology are autonomous. Explanations of cognitive phenomena need not necessarily refer to physical components. That is, identification of the component operations and demonstrating an organization of those operations that yields the phenomenon to be explained is sufficient for the phenomenon to be explained. The functionalist may also argue that there is a necessary tradeoff between the precision of an explanation and its generality (83–85).[8] Causal mechanical explanations might be more precise, but at the expense of generality. Functional explanations, which abstract over physical details, provide more general explanations that may apply to more diverse instantiations of the phenomenon. Both perspectives agree that functional analysis is useful, but they disagree on long-term explanatory goals.

Minimal Model Explanation

If we imagine causal mechanical explanation as one end of a continuum representing the tradeoff between precision and generality, minimal model explanation would be on the other extreme.[9] Recall that explanations answer why-questions. Batterman and Rice (86) distinguished between why-questions that ask why a phenomenon manifests in a particular situation and why-questions that ask why a phenomenon manifests generally or in a number of different circumstances. Minimal model explanations are concerned with the latter type while mechanistic explanations are concerned with the former.

Minimal model explanation employs mathematical abstraction techniques to delineate a set of physically distinct systems that demonstrate some shared behavior (86). Batterman writes,

> [E]xplanation of universal behavior involves the elucidation of principled reasons for bracketing or setting aside as "explanatory noise" many of the microscopic details that genuinely distinguish one system from another. In other words, it is a method for extracting just those features of systems, viewed macroscopically, that are stable under perturbation of their microscopic details. (87, p. 43)

This theory of explanation differs from the causal mechanical account in that the explanation need not share relevant features with the phenomenon to be explained. Instead, the explanation must abstract away from specific features of the phenomenon to enable wider generalization. It is also unlike functional explanation in that it does not assume a progressive decomposition into ever-simpler operations which eventually bottom out at some basic level of description. A minimal model explanation will not necessarily refer to component operations.

As an example of minimal model explanation in neuroscience, consider this analysis by Ross (88), who discusses the why-question, Why do neurons that differ drastically in the microstructural details all exhibit the same type of excitability? For background, there are at least three classes of neuron excitability (*Class I*, *Class II*, *Class III*). Excitability here refers to a neuron's relationship between input current and output firing rate. These classes represent large groups of physically distinct cells that display the same qualitative pattern of excitability. The canonical model approach to this question attempts to reduce the complexity of molecularly diverse neural systems to a single, abstracted model (the canonical model) that explains excitability in general, rather than for each distinct cell type. In this case, the canonical Fn8
ends up being a dynamical system, the Ermentrout–Kopell model, which shows how to transform all *Class I* cells into a canonical model of *Class I* excitability. This minimal model explanation describes the common reason why physically diverse *Class I* cells display the same excitability (88).

Dynamical and Computational Explanation in Neuroscience

Neuroscientists and cognitive scientists often build dynamical and computational models and may speak correspondingly of dynamical or computational explanation. These purported mathematical or computational explana- Fn9
tions are not associated with distinct philosophical theories of explanation. Instead, philosophers debate the degree to which such models provide satisfactory explanations according to existing theories of explanation (89).

There are many plausible interpretations. I group the strategies employed by philosophers of neuroscience into four main categories: *1*) One may conclude that the model does not in fact explain. The computational or mathematical model may be a descriptive or phenomenological model that captures or in some sense saves the phenomenon, without actually explaining its occurrence. *2*) One may conclude that

[8] A similar tradeoff can be found in the distinction between causal and unifying explanations described in Ref. 84 or in the tension between truth or accurate representation and the expansion of existing knowledge frameworks in Ref. 85.
[9] Functional explanation lies somewhere between the two, resembling causal mechanical explanation in that it is based on decomposition, but also resembling minimal model explanation in that it abstracts over physical details.

the model only explains if it meets the strict criteria for causal mechanical explanation, as codified in the 3M model described earlier (90, 91). Models that fail to meet these criteria may hope to do so in the future with further research. *3*) One may conclude that the model provides mechanistic explanations, but according to an adapted, more permissive conception of mechanistic explanation. This is the approach pursued by several philosophers who argue for the explanatory power of artificial neural network models in neuroscience (92–94). *4*) One may conclude that the model explains nonmechanistically (95). The abstract terms in the model, though not describing physical mechanisms, may still explain according to, for example, a functionalist or minimal model conception of explanation (or indeed nonmechanistic forms of causal explanation). For example, Chirimuuta (96) argues that explanations in computational neuroscience based on efficient coding amount to noncausal explanations. This strategy is also exemplified by the analysis of the canonical model of *Class I* excitability reviewed in the previous section (88). Elber-Dorozko's (97) proposal that computational models can provide noncausal yet still counterfactual explanations may also fit in this last category.

Considering these diverse proposals for how dynamical and computational models explain, it is not so surprising that I had trouble articulating what it would mean to "uncover the computational mechanisms underlying auditory perception" as a junior PhD student. Although this review does not provide one definitive answer to the question of how computational models explain, I am comforted by the knowledge of the options available to me to argue in favor of the explanatory power of a particular computational model. Personally, I'm currently most interested in how the value of AI models in neuroscience might be justified via the fourth strategy. I encourage my colleagues who also build models and hope to contribute toward explanatory goals to consider which of the strategies listed above is best suited to their work (or to propose alternative strategies, of course).

Aside on Statistical Explanation of Variance

Confusingly, the word *explain* is also used in the context of statistical hypothesis testing. The statistical measure *R*-squared (R^2) is the proportion of variance in one variable that is *explained* by another in a linear regression. This use of the word *explain* in statistical hypothesis testing is distinct from scientific explanation, but the two are sometimes not clearly distinguished in scientific writing. For example, consider this motivating statement for the Algonauts project, whose 2019 edition was dedicated to "Explaining the Human Visual Brain,"

> Currently, particular deep neural networks trained with the engineering goal to recognize objects in images do best in accounting for brain activity during visual object recognition (15, 98). However, a large portion of the signal measured in the brain remains unexplained. This is so because we do not have models that capture the mechanisms of the human brain well enough. Thus, what is needed are advances in computational modeling to better explain brain activity (99).

The authors allude to statistical explanation when discussing unexplained signals, while talk of capturing neural mechanisms hints to scientific explanation. In reality, the Algonauts project is about evaluating models based on their ability to predict (in the machine learning sense) neural activity. When they lament that a "large portion of the signal measured in the brain remains unexplained," they invoke the notion of explained variance. Rather than trying to develop a scientific explanation for a phenomenon of interest, they are concerned with statistically explaining, or in this case, being able to predict, the variance in the collected data —variance which may or may not be causally related to any number of different neural or cognitive phenomena.

The goal to explain as much variance as possible or to predict as accurately as possible expresses a desire for completeness. Philosophers of scientific explanation warn against overcompleteness.

> It is important to realize that we cannot aspire to explain particular phenomena in their full particularity. In explanations of particular phenomena, the explanation-seeking why-question—suitably clarified and reformulated if necessary—should indicate [only] those aspects of the phenomena for which an explanation is sought. (63, p. 273–274)

The project of collecting large-scale neural datasets and building models that explain as much variance as possible in that data is one of mere description rather than explanation. Descriptive science is unambiguously crucially important to scientific progress. Before we can hope to explain, we must first characterize the phenomenon to be explained[10] (62). Fn10
However, the distinction between explanation and mere description is still important. We can distinguish between understanding-that, understanding-how, and understanding-why a particular phenomenon occurs (100). Surely we need to understand-that before we can even begin to understand-why, but only answers to understand-why questions are explanatory (64, p. 96). Specific why-questions may eventually be motivated by descriptive characterizations, but only if we do not mistake them for explanations prematurely.

SCIENTIFIC EXPLANATION AT THE INTERSECTION OF NEUROSCIENCE AND AI

Phenomena-Specific Theories of Explanation to Unify Neuroscience and AI

One development in the recent history of philosophy of science is a trend toward domain-specific theories. Previous efforts in the mid-twentieth century sought to develop a universal logic of science that would apply to all fields. This endeavor encountered several challenges. For one, the

[10]This acknowledgment is reflected in Craver's view which lists characterization of the phenomenon to be explained as the first aspect of mechanistic explanation (62).

norms of explanation appeared to be domain-specific; explanation in biology is not the same as explanation in physics. So now instead we have subareas of philosophy of science dedicated to philosophy of biology, philosophy of psychology, and even philosophy of neuroscience. This has led to the development of theories of explanation specific to those fields.

However, given the highly interdisciplinary nature of neuroscience, I am skeptical about the feasibility of a single theory of explanation to account for all explanation in neuroscience. Neuroscience is not clearly separated from its sister sciences: biology, physics, psychology, AI, etc. Thus, the search for a universal theory of explanation in neuroscience may be as ill-fated as the original quest for a theory of explanation for all science. I interpret the original motivation for field-specific theories of explanation as essentially that similar phenomena ought to be explained similarly, assuming that phenomena within a branch of science will be more similar than phenomena from different branches. I would argue that this assumption does not hold for neuroscience where, for example, one phenomenon might be closer to biophysics and another closer to psychology than the two are to each other. Therefore, rather than organizing our theories of explanation around objects of study (in this case, the brain) or the departmental silos of our academic institutions, I suggest we organize our theories around classes of similar phenomena, regardless of the specific scientific discipline the phenomena "belong" to.

Such an organization can be especially unifying at the intersection of neuroscience and AI where artificial systems are designed to perform feats of human and animal cognition. This behavioral mimicry brings into alignment the scientific goals of AI and much of computational neuroscience, which both seek to identify the principles of neural computation that underlie cognitive abilities. The AI system and the biological system it models both demonstrate, to some extent, the same phenomenon to be explained. Thus, I propose we consider theories of explanation that are specific to the classes of phenomena we associate with animal intelligence, such as learning and decision making, irrespective of whether they manifest in biological or artificial systems.

A related analysis is presented by Stinson (101). Stinson observes that much of the skepticism about neural network or connectionist models in neuroscience and cognitive science is concerned with their dissimilarity to the target they are said to model, e.g., lack of biological plausibility or omission of physiological details. To understand the usefulness of connectionist models, despite their apparent dissimilarity, she proposes that both the model and target are instances of the same *kind*.[11] In contrast to other accounts of modeling, which say that a model ought to be similar to its target system, Stinson's analysis includes an additional step of associating the model first with a kind, then that kind with the target. Rather than a direct relationship between a model and its target, their belonging to the same kind mediates an indirect relationship. This analysis accounts well for connectionist models as idealized models of cognitive mechanisms. The lack of physiological detail or biological plausibility in connectionist models is not a mistake—nor does it make them necessarily less useful—because complete and faithful imitation is not the goal. Useful models will often be minimal, focusing on a particular aspect or phenomenon of the target system, not the entire system as a whole. In short, connectionist models and their targets need not be especially similar to be doing the same thing (101).

One consequence of this phenomena-specific view is that it should be possible to apply the same theory of explanation to both artificial and natural instantiations of similar phenomena. If what constitutes a valid scientific explanation is dependent on the phenomenon to be explained rather than on the field or object of study, then, to the extent that an artificial and biological system demonstrate the same phenomenon, what constitutes a valid explanation of that phenomenon should be the same for both systems, even if the content of the explanations differ. For example, Geirhos et al. (102) showed that convolutional neural networks trained the ImageNet dataset are biased to recognize texture rather than shape, whereas humans privilege shape over texture when making decisions about object category. Analyzing which visual features an agent uses to make decisions about an image reflects a functionalist approach where the detection of individual features are the component operations that combine to yield the decision—the phenomenon to be explained. Although the two explanations differ in content (one prefers texture, the other shape), the same approach to explanation (decomposition into component operations) is applied to both artificial and biological intelligence.

Thus, when evaluating theories of scientific explanation for why-questions about phenomena that occur in both artificial and biological systems, we can insist that the theory of explanation be appropriate for both manifestations. This may lead us to eliminate some candidate theories that appear appropriate in one context but not the other. For example, the strictest version of the causal mechanical theory of explanation seems ill-suited to explain many behaviors of AI systems. This assessment is consistent with the various proposals for modified theories of mechanical explanation to account for the explanatory power of neural network models.

Explanations in Neuro-AI Answer Why-Questions about Phenomena That Are Common to Both Artificial and Natural Intelligence

Within this phenomenon-specific framework, it becomes very important to clearly identify and delineate the phenomena to be explained. How a scientist conceptualizes the phenomenon to be explained may bias them toward one form of explanation or another. This is especially apparent in the cognitive sciences where there are many different perspectives on the nature of mind and cognition. Is cognition computation? Are cognitive agents embodied dynamical systems? Is the brain a control system? If explanations are phenomena-specific, we cannot completely separate the ontological question (What is cognition?) from the epistemological question (How do we explain cognition?). Thus, a commitment to a particular

Fn11

[11]The notion of kind has a long history in philosophy. For our purposes, we can think of a kind as simply a prototype or template for some scientific category.

theory of explanation in neuroscience may also suggest a related commitment to a theory of cognition. For example, the information processing view of vision, as exemplified by pioneers like Marr (103), may bias researchers toward functional explanations. Alternatively, a radical embodied perspective, like that espoused by Chemero (104), might bias researchers toward dynamical explanations and away from explanations that rely on intentional representations.

We may want to allow that there is not one single correct answer—no one correct way to conceptualize the mind, no one correct answer to how to explain mind, and no one correct approach to studying mind. Embracing any or all of these would amount to a kind of scientific pluralism[12] (105, 106). Several contemporary neuroscientists and philosophers of science have espoused some version of pluralism (83, 95, 96, 107–109), but traditional conceptions of explanation may not easily accommodate explanatory pluralism. Traditionally, we say that an explanation is a description of both the thing to be explained (some phenomenon) and the thing that explains (e.g., a mechanism or a program). This can suggest a one-to-one mapping between a phenomenon and it's explanation, limiting the ability to describe more nuanced one-to-many or many-to-many relationships. A useful strategy for making sense of explanatory pluralism may be to focus on the specific why-questions to be answered. This approach is exemplified by Krakauer et al. (108), who proposed that a plurality of explanations of a single phenomenon may answer questions related to, for example, Tinbergen's four questions (causation, evolution, survival, ontogeny) (110) or Marr's levels (implementation, representation and algorithm, computation) (103). As such, it may be more parsimonious to focus on the conception of an explanation as an answer to a why-question, where there are many why-questions to be asked about a single phenomenon and each why-question can have variable scope.

The scope of the why-question to be answered will determine what features of the phenomenon will be abstracted over in the explanation. Many why-questions at the intersection of neuroscience and AI are similar to the why-questions that the minimal model theory of explanation are said to address. The canonical model approach as discussed by Ross (88), "explains why physically distinct neural systems all share the same behavior by showing that principled mathematical abstraction techniques—which preserve qualitative behavior—can be used to reduce all models of these distinct systems to the same canonical model" (p. 15). However, rather than the two distinct categories of why-questions (why-particular versus why-general), as discussed by Batterman, it seems to me that all why-questions require a clause of scope, where that scope can vary continuously from more to less specific. For example, why-questions about why certain behaviors are exhibited by a particular artificial neural architecture have a scope that includes all instantiations of that architecture, but not other architectures and not human brains. Consider an AI system with human-level ability to recognize faces. A canonical model may explain why the AI system and a human demonstrate (or do not demonstrate) the same behavior (111, 112). We can also ask why-questions about learning in distributed networks, the answers to which would hold for some class of networks, regardless of whether they were implemented in cells or silicon. On the other hand, some why-questions will be about a specific manifestation of a phenomenon (e.g., in human brains or deep neural networks (DNNs) or brains of a particular clinical population). In these cases, the why-question and its answer, appropriately stated, may not abstract over the relevant features that define the particular manifestation in question. Thus, perhaps the minimal model theory can be applied to seemingly more why-particular questions as well, not only why-general questions.

To be precise, I propose that explanations answer why-questions that include both a description of the phenomenon to be explained and the scope in which the answer should apply. An alternative view might say that the relevant scope is part of the definition of the phenomenon itself. I prefer to keep them separate because it allows us to discuss why-questions that are concerned with phenomena that are exhibited in both biological and artificial systems, while allowing the scope to be an additional, separate variable. For example, Leavitt and Morcos (113) recently posed the question, Why does class selectivity emerge in deep neural networks trained on classification tasks? The phenomenon, the emergence of class selectivity, also occurs in animal brains, therefore we can additionally ask, Why does class selectivity emerge in rodent brains? or Why does class selectivity emerge in human brains? or Why does class selectivity emerge in both rodent and human brains? or Why does class selectivity emerge in both DNNs and human brains? In all cases, the why-question is asking what is the common reason that this phenomenon (emergence of class selectivity) occurs in some set of observations, where the scope of that set is larger (e.g., all animals) or smaller (e.g., DNNs of a particular architecture). I propose that the why-questions at the intersection of neuroscience and AI are those about phenomena that occur (to some extent) in both artificial and biological intelligence, where the scope may or may not include both. This is a relatively broad definition as it includes both AI research that does not appear to care about the brain and brain research that does not appear to care about AI, depending only on the phenomenon of study, not the organism in which it occurs.

The Deep Learning Framework for Neuroscience Proposes Alternative Why Questions

I interpret the deep learning approach to neuroscience, as described in Richards et al. (46) but also defended in Lillicrap and Kording (65) and Marblestone et al. (38), as primarily an invitation to consider particular why-questions. This framework focuses on how architectures, cost functions, and learning algorithms produce intelligent behavior like learning and decision making. Contrary to the title "What does it mean to understand a neural network?" Lillicrap and Kording do not provide a recipe for achieving scientific understanding. Instead, their conclusions are really about which scientific questions are most readily answerable.

I find the distinction between etiological and constitutive mechanisms summarized in THEORIES OF SCIENTIFIC EXPLANATION useful here. I agree with Craver that most contemporary references to mechanisms in neuroscience

[12]For more on scientific pluralism, see Refs. 105 or 106.

reflect a commitment to constitutive explanation. Part of the resistance to or confusion about the deep learning framework for neuroscience is because this framework is concerned primarily with etiological mechanisms rather than constitutive ones. Within this framework, learning and model selection procedures are viewed as structural, etiological mechanisms that establish the conditions such that a trigger, such as an image, yields the phenomenon to be explained, for example, the recognition of an object in an image. For comparison, a constitutive story about that same recognition would amount to describing the component parts of the network (the units and weights) and how they interact (after training) to realize the act of recognition. In the constitutive story, the recognition and learning to recognize might be two separate phenomena to be explained with their own constitutive mechanisms. In the etiological story, the learning is one of potentially many structural mechanisms that together set up the conditions such that the recognition process, a triggering mechanism, can occur.

Etiological mechanisms are commonly described in the social and psychological sciences. For example, a race scholar may describe the mechanisms by which racist policies produce racial inequality (114), or a clinical psychologist may describe the mechanisms by which some therapy produces behavioral change in a patient. Neuroscience, even when studying phenomena like learning and memory, often seeks constitutive stories about how the physical components of brains (cell types, cell parts, neurotransmitters, brain regions, circuits) and their activities are organized to realize the learning or memory to be explained. Machine learning research, on the other hand, has largely focused on etiological stories. The proposal for a deep learning approach to neuroscience invites neuroscientists to consider the kinds of etiological stories that have been useful in deep learning research.

The Goal of Explainable AI Is Not the Same as the Goal of Scientific Explanation of AI

In machine learning and AI spheres, the word *explanation* is most likely to arise in the context of explainable of interpretable AI, which is commonly posed as an applied research, human-computer-interaction problem. AI systems need to be "explainable" or "interpretable" to their users for many applications, especially in cases where AI systems are making decisions of consequence, such as medical diagnosis. For such technologies to be used, a user needs to have a degree of trust in the AI's predictions. This trust can be established by making the artificial decision process more transparent, either by constructing models that are transparent by design or by developing tools to "open the black box" of more opaque systems (115). For example, linear models are often said to be interpretable because their weights describe linear relationships between variables (e.g., tobacco use and cancer risk) that are immediately understandable by a user [However, note that Lipton (116) and Haufe et al. (117) have questioned this intuition that linear models are inherently more interpretable].

Viewed as an applied research problem, the connection between explainable AI and scientific explanation is not immediately obvious. Scientific explanation is not sought as a prerequisite for the adoption of new technologies. It is also not clear that scientific explanation would achieve the goal of explainable AI. Sometimes the issue may be one of justification rather than explanation, where the AI system needs to justify its decision or recommendation according to the same norms that a human would adhere to when doing same job. For example, in legal or political applications, there may not even be a true or correct rationalization. All that the AI system can hope for is a justification of a policy recommendation or sentencing that satisfies the same norms that a politician or a judge would be expected to follow.

In addition, scientific explanations of AI behavior may be intelligible and useful to AI scientists, but not to the users of AI technology, who may not possess the necessary skills and background knowledge. Buckner (118) expands on this distinction between justificatory and explanatory rationalizations in the context of explainable AI, pointing out that human subjects often do not prefer the most causally accurate rationalization. He similarly argues that the best causal explanation for a network's behavior may not cite factors that are intelligible to human users and that the most compelling justifications may not cite causal factors. Thus, the goals of explainable AI should not be equated with the goal of scientific explanation of AI. Scientific explanations may not achieve the explainable AI goal and accepted rationalizations for the behavior of AI systems may not meet the norms for scientific explanation.

Other related areas of research in machine learning and AI have scientific explanation as a primary goal. Traditionally, machine learning theory may have been home to the more scientific aspects of machine learning, albeit for a very limited set of phenomena related to the practical use of machine learning systems, e.g., their generalization performance, convergence rate, or sample efficiency. The advancement of deep learning, where existing theory failed to account for phenomena like the generalization performance of massively over-parameterized neural networks, has led to increased interest in empirical methods to help fill in the gaps of the existing analytical theory.

There is a growing subdiscipline in machine learning, sometimes referred to as "science of deep learning" (119) or "understanding deep learning," committed to scientific explanation and understanding of AI systems. The ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena solicited "contributions that view the behavior of deep nets as natural phenomena to be studied with methods inspired from the natural sciences like physics, astronomy, and biology," The National Academy of Sciences hosted a colloquium in 2019 on "The Science of Deep Learning" asked (among other questions) "Can experimental techniques be used to study the nature of artificial deep neural networks?" (120). Similar research has also been referred to as "artificial neuroscience" (121) or "synthetic neurophysiology" (35), to highlight the similarity of both methodologies and goals. Methods like ablations (122), single neuron selective-response characterization (123), saliency maps (124), similarity analysis (125), and dynamical systems modeling (126) are similar to those employed in neuroscience to study biological neural processing[13] (127). A science of deep Fn13

[13]see Ref. 127 for a critical overview of some of these methods.

learning is not just about building better AI systems (which is fundamentally an engineering goal, not a scientific one) and the phenomena of interest extend beyond task performance, convergence rates, and generalization bounds to include things like the shape of loss landscape, learning trajectories, and adversarial examples.

REAPING THE BENEFITS OF AN INTERDISCIPLINARY COGNITIVE COMPUTATIONAL NEUROSCIENCE

Up to now, I have discussed avenues for the unification of scientific goals in neuroscience and AI. However, this should not be interpreted as proposing a convergence to a single scientific approach or methodology. The scientists that make up cognitive computational neuroscience come from a large variety of fields: cognitive science, psychology, neuroscience, and AI to name a few. Better understanding of our shared long-term goals does not negate the diversity of our near-term goals. Proposing new why-questions does not invalidate those already posed. When we understand science as a cooperative social process, this diversity can be embraced as a strength.

Feminist philosophers of science have analyzed science as a social rather than individual enterprise and have focused their theorizing on the process of doing science in a scientific community, rather than solely on the products of science (theories, explanations, etc.). An important claim in this space is that a scientific community's degree of objectivity will be related to its ability to self-scrutinize (128, 129). Heterogenous science communities, it is argued, have greater potential for objectivity due to their diverse background assumptions, which better enable them to recognize, scrutinize, and modify such assumptions (85). According to Harding's influential standpoint theory, knowers are "situated," their vantage points influencing what they can know about themselves and the world (130). This situatedness justifies the call for "more diverse science communities, communities such that the methodological and metaphysical assumptions functioning as evidence for specific research questions and claims would be subjected to a broader range of scrutiny" (128, p. 323). Although perhaps uncomfortable to many scientists who would like to think of themselves as objective, unbiased, rational truth-seekers, this perspective does not necessarily reduce to antirealism. It suggests that we access the truth in cooperation and in concert, not in isolation, and that who we cooperate with matters.

A similar claim has been investigated empirically. Devezer et al. (131) simulated scientific discovery in a model-centric framework to characterize the relationship between several attributes of scientific communities and the success of their research programs. The authors found that innovative research sped up the discovery of scientific truth by facilitating the exploration of model space. They also observed that epistemic diversity (here, the use of several modeling approaches) optimized scientific discovery by protecting against ineffective research strategies.

What is cognitive computational neuroscience's capacity for self-reflection and self-scrutiny? It ought to be great, given the interdisciplinary nature of our field and the plethora of background assumptions and methodologies that entails. The recent resurgence of AI and machine learning models in neuroscience may have spurred a period of increased self-scrutiny precisely because of the differing background assumptions of these fields. This is a good thing! However, care will be required to reap the benefits of our heterogeneity.

We might be guided by some of the recommendations in philosophical work on the social processes by which science progresses. For example, Longino (85) proposed four criteria that must be met for a scientific community to experience the transformative aspects of critical discourse:

> *1*) there must be recognized avenues for the criticism of evidence, of methods, and of assumptions and reasoning; *2*) there must exist shared standards that critics can invoke; *3*) the community as a whole must be responsive to such criticism; *4*) intellectual authority must be shared equally among qualified practitioners. (85, p. 76)

Criteria 1–3 tell us that criticism alone is not enough; the criticism must refer to some shared goals or values, it must be valued, and the scientific community must be sensitive to it. *Criterion 4* is meant to protect against the scenario where a particular set of assumptions is suppressed due to the lack of political power of its adherents. An example of the violation of *Criterion 4* is the historical exclusion of certain gender, racial, and ethnic groups from science and engineering.

How can we interpret these recommendations in the context of contemporary cognitive computational neuroscience? Working together as an interdisciplinary community will mean embracing many why-questions with varying scope, which may have multiple answers that could take generations to discover. To make progress toward these challenging questions, we must resist the isolationist temptation to ignore conflicting perspectives or to give epistemic authority to one field over another. Further clarification of the specific targets and forms of our explanations will better facilitate interdisciplinary synthesis and scrutiny. This perspective also implies that addressing the exclusion of members of certain gender, racial, ethnic, and other identity groups in our field is not just a moral imperative, but also an epistemic one. Addressing these inequalities and supporting constructive criticism will better facilitate our cooperation with current and future scientists.

CONCLUSIONS

I hope that I have successfully demonstrated some of the areas where philosophy of science can inform scientific practice at the intersection of neuroscience and AI. Philosophy of science can help us enumerate, define, and choose among possible scientific goals. It can describe the various ways that phenomena have been explained and translate those descriptions into normative criteria. It can help us to carve up a topic of study into specific phenomena to be explained and scientific questions to be posed. It helps us to understand the relationship between a model and the target system it is said to model. It can help to situate individual scientific activities into broader, long-term scientific goals. It can help us to conceptualize our position relative to other

scientists in our community and to imagine productive cooperation among diverse and disputing researchers.

Unsatisfactorily to some, this body of philosophy does not prescribe the actions that a particular scientist should take. Explanations are not typically the product of individual experiments or even individual research groups. Achieving explanatory understanding of complex cognitive and neural phenomena will require the sustained coordination of a diverse scientific community.

Along the way to some eventual explanatory understanding, we will formulate and answer many other important questions, but our progress will be stalled if we mistake these milestones for the finish line. Having an intelligible theory is not sufficient for explanatory understanding. Statistically explaining all the variance in our data is not explanatory understanding. An established causal relationship is typically not by itself going to provide the desired explanatory understanding. Any result of a single experiment is not explanatory understanding. On the other hand, our progress will be expedited if we have a common road map. That common road map will not specify a particular research approach, of which we surely need many. But it may have a common conception of the different types of scientific questions our community wants to ask, what the various answers might look like, and how these answers may be integrated to formulate new questions.

Toward the goal of a common road map, I proposed that why-questions at the intersection of neuroscience and AI are concerned with phenomena that occur to some degree in both natural and artificial systems, posed at a particular scope. The scope will place the question along a continuum from why-general to why-specific, which may suggest candidate forms of explanation, e.g., minimal model explanation for why-general. I interpreted the deep learning approach to neuroscience as proposing answerable and previously underexplored why-questions about the etiological mechanisms that produce intelligent behavior. I also clarified that the goal of explainable AI is not necessarily the same as the goal for scientific explanations of the behavior of AI systems. Lastly, I discussed why we should value scientific criticism, methodological pluralism, and diverse perspectives if we are to reap the benefits of our interdisciplinarity. Ultimately, innovative and diverse approaches in an epistemically humble research community will better lead us toward our goals.

ACKNOWLEDGMENTS

I would like to thank Rosa Cao, Chris Summerfield, Marc Schönwiesner, Christoph Daube, Harris Kaplan, Catherine Stinson, Raphael Gerraty, and the NYU NeuroPhil Journal Club for helpful comments and feedback on earlier drafts.

GRANTS

This work was funded in part by Fonds de recherche du Québec—Nature et technologies (FRQNT).

ACKNOWLEDGMENTS

I would like to thank Rosa Cao, Chris Summerfield, Marc Schönwiesner, Christoph Daube, Harris Kaplan, Catherine Stinson, Raphael Gerraty, and the NYU NeuroPhil Journal Club for helpful comments and feedback on earlier drafts.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author.

AUTHOR CONTRIBUTIONS

J.T. drafted manuscript; edited and revised manuscript; approved final version of manuscript.

REFERENCES

1. **Naselaris T**, **Bassett DS**, **Fletcher AK**, **Kording K**, **Kriegeskorte N**, **Nienborg H**, **Poldrack RA**, **Shohamy D**, **Kay K.** Cognitive computational neuroscience: a new conference for an emerging discipline. *Trends Cogn Sci* 22: 365–367, 2018. doi:10.1016/j.tics.2018.02.008.
2. **Krizhevsky A**, **Sutskever I**, **Hinton GE.** ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, edited by Pereira F, Burges CJC, Bottou L, Weinberger KQ. NeurIPS Proceedings, 2012.
3. **Lindsay G.** *Models of the Mind: How Physics, Engineering and Mathematics Have Shaped Our Understanding of the Brain*. London: Bloomsbury Publishing, 2021.
4. **Yamins DLK**, **Hong H**, **Cadieu C**, **DiCarlo JJ.** Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. In: *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, edited by Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ. NeurIPS Proceedings, 2013.
5. **Stansbury DE.** *Modeling Neural Representation Using Statistical Features of Natural Scenes*. Berkeley, CA: University of California, 2014.
6. **Agrawal P**, **Stansbury D**, **Malik J**, **Gallant JL.** Pixels to voxels: modeling visual representation in the human brain (Preprint). *arXiv* 1407.5104, 2014. https://arxiv.org/abs/1407.5104.
7. **Cadieu CF**, **Hong H**, **Yamins DLK**, **Pinto N**, **Ardila D**, **Solomon EA**, **Majaj NJ**, **DiCarlo JJ.** Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Comput Biol* 10: e1003963, 2014. doi:10.1371/journal.pcbi.1003963.
8. **Yamins DLK**, **Hong H**, **Cadieu CF**, **Solomon EA**, **Seibert D**, **DiCarlo JJ.** Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci USA* 111: 8619–8624, 2014. doi:10.1073/pnas.1403112111.
9. **Güçlü U**, **van Gerven MAJ.** Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J Neurosci* 35: 10005–10014, 2015. doi:10.1523/JNEUROSCI.5023-14.2015.
10. **Dubey A**, **Jayadeva, Agarwal S.** Examining representational similarity in ConvNets and the primate visual cortex (Preprint). *arXiv* 1609.03529, 2016. https://arxiv.org/abs/1609.03529.
11. **Cichy RM**, **Khosla A**, **Pantazis D**, **Torralba A**, **Oliva A.** Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci Rep* 6: 27755, 2016. doi:10.1038/srep27755.
12. **Storrs K**, **Mehrer J**, **Walther A**, **Kriegeskorte N.** Architecture matters: how well neural networks explain IT representation does not depend on depth and performance alone. *Conference on Cognitive Computational Neuroscience (CCN)*, 2017.
13. **Spoerer CJ**, **McClure P**, **Kriegeskorte N.** Recurrent convolutional neural networks: a better model of biological object recognition. *Front Psychol* 8: 1551, 2017. doi:10.3389/fpsyg.2017.01551.
14. **Jozwik KM**, **Kriegeskorte N**, **Storrs KR**, **Mur M.** Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Front Psychol* 8: 1726, 2017. doi:10.3389/fpsyg.2017.01726.
15. **Bashivan P**, **Kar K**, **DiCarlo JJ.** Neural population control via deep image synthesis. *Science* 364: eaav9436, 2019. doi:10.1126/science.aav9436.

16. **Kietzmann TC**, **Spoerer CJ**, **Sörensen LKA**, **Cichy RM**, **Hauk O**, **Kriegeskorte N.** Recurrence is required to capture the representational dynamics of the human visual system. *Proc Natl Acad Sci USA* 2019;116:21854–21863. doi:10.1073/pnas.1905544116.
17. **Kar K**, **Kubilius J**, **Schmidt K**, **Issa EB**, **DiCarlo JJ.** Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat Neurosci* 22: 974–983, 2019. doi:10.1038/s41593-019-0392-5.
18. **Konkle T**, **Alvarez GA.** Instance-level contrastive learning yields human brain-like representation without category-supervision (Preprint). *bioRxiv*, 2020. doi:10.1101/2020.06.15.153247.
19. **Zhuang C**, **Yan S**, **Nayebi A**, **Schrimpf M**, **Frank MC**, **DiCarlo JJ.** Unsupervised neural network models of the ventral visual stream (Preprint). *bioRxiv*, 2020. doi:10.1101/2020.06.16.155556.
20. **Kell AJE**, **Yamins DLK**, **Shook EN**, **Norman-Haignere SV**, **McDermott JH.** A task-optimized neural network replicates human auditory predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* 98: 630–644.e16, 2018. doi:10.1016/j.neuron.2018.03.044.
21. **Güçlü U**, **Thielen J**, **Hanke M**, **van Gerven MAJ.** Brains on beats. In: *Advances in Neural Information Processing Systems 29*, edited by Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R. NeurIPS Proceedings, 2016**.**
22. **Cadena SA**, **Sinz Fh Muhammad T**, **Froudarakis E**, **Cobos E**, **Walke EY**, **Reimer J**, **Bethge M**, **Tolias A**, **Ecker AS.** How well do deep neural networks trained on object recognition characterize the mouse visual system? *Real Neurons & Hidden Units NeurIPS*, 2019.
23. **Nayebi A**, **Zhuang C**, **Norcia AM**, **Kong NCL**, **Gardner JL**, **Yamins DLK.** Unsupervised models of mouse visual cortex (Preprint). *bioRxiv* 2021.06.16.448730, 2021. doi:10.1101/2021.06.16.448730.
24. **Bakhtiari S**, **Mineault P**, **Lillicrap T**, **Pack CC**, **Richards BA.** The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning (Preprint). *bioRxiv* 2021.06.18.448989, 2021. doi:10.1101/2021.06.18.448989.
25. **Guerguiev J**, **Lillicrap TP**, **Richards BA.** Towards deep learning with segregated dendrites. *eLife* 6: e22901, 2017. doi:10.7554/eLife.22901.
26. **Wang JX**, **Kurth-Nelson Z**, **Kumaran D**, **Tirumala D**, **Soyer H**, **Leibo JZ**, **Hassabis D**, **Botvinick M.** Prefrontal cortex as a meta-reinforcement learning system. *Nat Neurosci* 21: 860–868, 2018. doi:10.1038/s41593-018-0147-8.
27. **Richards BA**, **Lillicrap TP.** Dendritic solutions to the credit assignment problem. *Curr Opin Neurobiol* 54: 28–36, 2019. doi:10.1016/j.conb.2018.08.003.
28. **Russin J**, **O'Reilly RC**, **Bengio Y.** Deep learning needs a prefrontal cortex. *Bridging AI and Cognitive Science (ICLR 2020)*, 2020. https://baicsworkshop.github.io/pdf/BAICS_10.pdf.
29. **Schaeffer R**, **Khona M**, **Meshulam L**, **Fiete IR.** Reverse-engineering recurrent neural network solutions to a hierarchical inference task for mice (Preprint). *bioRxiv* 2020.06.09.142745, 2020. doi:10.1101/2020.06.09.142745.
30. **Gu K**, **Greydanus S**, **Metz L**, **Maheswaranathan N**, **Sohl-Dickstein J.** Meta-learning biologically plausible semi-supervised update rules (Preprint). *bioRxiv* 2019.12.30.891184, 2019. doi:10.1101/2019.12.30.891184.
31. **Chung S**, **Abbott LF.** Neural population geometry: an approach for understanding biological and artificial neural networks (Preprint). *arXiv* 210407059, 2021. https://arxiv.org/abs/2104.07059.
32. **Flesch T**, **Balaguer J**, **Dekker R**, **Nili H**, **Summerfield C.** Comparing continual task learning in minds and machines. *Proc Natl Acad Sci USA* 2018; 115: E10313–E10322. doi:10.1073/pnas.1800755115.
33. **Bernardi S**, **Benna MK**, **Rigotti M**, **Munuera J**, **Fusi S**, **Salzman CD.** The geometry of abstraction in hippocampus and prefrontal cortex. *Cell* 183: 954–967, 2020. doi:10.1016/j.cell.2020.09.031.
34. **Hinton GE.** Machine learning for neuroscience. *Neural Syst Circuits* 1: 12, 2011. doi:10.1186/2042-1001-1-12.
35. **Kriegeskorte N.** Deep neural networks: a new framework for modelling biological vision and brain information processing. *Annu Rev Vis Sci* 1: 417–446, 2015. doi:10.1146/annurev-vision-082114-035447.
36. **Thompson JAF**, **Bengio Y**, **Formisano E**, **Schönwiesner M.** How can deep learning advance computational modeling of sensory information processing? Neural Information Processing Systems workshop on Representation Learning in Artificial and Biological Neural Networks. 2016.
37. **Yamins DLK**, **DiCarlo JJ.** Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci* 19: 356–365, 2016. doi:10.1038/nn.4244.
38. **Marblestone AH**, **Wayne G**, **Kording KP.** Towards an integration of deep learning and neuroscience. *Front Comput Neurosci* 10: 94, 2016. doi:10.3389/fncom.2016.00094.
39. **van Gerven M**, **Bohte S.** Editorial: Artificial neural networks as models of neural information processing. *Front Comput Neurosci* 11: 114, 2017. doi:10.3389/fncom.2017.00114.
40. **Kubilius J.** Predict, then simplify. *NeuroImage* 180: 110–111, 2018. doi:10.1016/j.neuroimage.2017.12.006.
41. **Scholte HS.** Fantastic DNimals and where to find them. *NeuroImage* 180: 112–113, 2018. doi:10.1016/j.neuroimage.2017.12.077.
42. **Turner BM**, **Miletić S**, **Forstmann BU.** Outlook on deep neural networks in computational cognitive neuroscience. *NeuroImage* 180: 117–118, 2018. doi:10.1016/j.neuroimage.2017.12.078.
43. **Tripp B.** A deeper understanding of the brain. *NeuroImage* 180: 114–116, 2018. doi:10.1016/j.neuroimage.2017.12.079.
44. **Kay KN.** Principles for models of neural information processing. *NeuroImage* 180: 101–109, 2018. doi:10.1016/j.neuroimage.2017.08.016.
45. **Ganguli S.** The Intertwined Quest for Understanding Biological Intelligence and Creating Artificial Intelligence. 2018. https://neuroscience.stanford.edu/news/intertwined-quest-understanding-biological-intelligence-and-creating-artificial-intelligence [2020 April 30].
46. **Richards BA**, **Lillicrap TP**, **Beaudoin P**, **Bengio Y**, **Bogacz R**, **Christensen A**, **Clopath C**, **Costa RP**, **de Berker A**, **Ganguli S**, **Gillon CJ**, **Hafner D**, **Kepecs A**, **Kriegeskorte N**, **Latham P**, **Lindsay GW**, **Miller KD**, **Naud R**, **Pack CC**, **Poirazi P**, **Roelfsema P**, **Sacramento J**, **Saxe A**, **Scellier B**, **Schapiro AC**, **Senn W**, **Wayne G**, **Yamins D**, **Zenke F**, **Zylberberg J**, **Therien D**, **Kording KP.** A deep learning framework for neuroscience. *Nat Neurosci* 22: 1761–1770, 2019. doi:10.1038/s41593-019-0520-2.
47. **Zador AM.** A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat Commun* 10: 3770, 2019. doi:10.1038/s41467-019-11786-6.
48. **Cichy RM**, **Kaiser D.** Deep neural networks as scientific models. *Trends Cogn Sci* 23: 305–317, 2019. doi:10.1016/j.tics.2019.01.009.
49. **Kell AJ**, **McDermott JH.** Deep neural network models of sensory systems: windows onto the role of task constraints. *Curr Opin Neurobiol* 55: 121–132, 2019. doi:10.1016/j.conb.2019.02.003.
50. **Kietzmann TC**, **McClure P**, **Kriegeskorte N.** Deep neural networks in computational neuroscience (Preprint). *bioRxiv* 2019. doi:10.1101/133504.
51. **Barrett DG**, **Morcos AS**, **Macke JH.** Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Curr Opin Neurobiol* 55: 55–64, 2019. doi:10.1016/j.conb.2019.01.007.
52. **Storrs KR**, **Kriegeskorte N.** Deep learning for cognitive neuroscience. In: *The Cognitive Neurosciences* (6th ed.), edited by Poeppel D, Mangun GR, Gazzaniga MS. Cambridge, MA: MIT Press, 2020, p**.** 707–716.
53. **Hasson U**, **Nastase SA**, **Goldstein A.** Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron* 105: 416–434, 2020. doi:10.1016/j.neuron.2019.12.002.
54. **Saxe A**, **Nelli S**, **Summerfield C.** If deep learning is the answer, then what is the question? *Nat Rev Neurosci* 22: 55–67, 2021.
55. **Machamer P**, **Silberstein M** (Editors)**.** *The Blackwell Guide to the Philosophy of Science*. Malden, MA: Blackwell Publishers, 2002.
56. **Woodward J.** Scientific explanation. In: *The Stanford Encyclopedia of Philosophy* (Fall 2017 ed.), edited by Zalta EN. Metaphysics Research Lab, Stanford University, 2017**.** https://plato.stanford.edu/archives/fall2017/entries/scientific-explanation/.
57. **Salmon WC.** *Four Decades of Scientific Explanation*. Minneapolis, MN: University of Minnesota Press, 1989.
58. **Santoro R**, **Moerel M**, **De Martino F**, **Goebel R**, **Ugurbil K**, **Yacoub E**, **Formisano E.** Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Comput Biol* 10: e1003412, 2014. doi:10.1371/journal.pcbi.1003412.
59. **Rosenblueth A**, **Wiener N.** The role of models in science. *Philos Sci* 12: 316–321, 1945. doi:10.1086/286874.
60. **Grimm S.** Understanding. In: *The Routledge Companion to Epistemology*, edited by Berneker S, Pritchard D. New York: Routledge, 2011, p**.** 84–94.
61. **Hempel CG.** *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: The Free Press, 1965.

62. **Craver CF.** *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford, UK: Clarendon Press, 2007.
63. **Salmon WC.** *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press, 1984.
64. **De Regt HW.** *Understanding Scientific Understanding*. New York: Oxford University Press, 2017.
65. **Lillicrap TP**, **Kording KP.** What does it mean to understand a neural network? (Preprint). *arXiv* 1907.06374, 2019. https://arxiv.org/abs/1907.06374.
66. **Kelly KT.** Simplicity, truth, and the unending game of science. In: *Infinite Games: Foundations of the Formal Sciences V*, edited by Bold S, Lowe B, Rasch T, van Benthem J**.** New York: College Publications, 2007, p. 223–270.
67. **Woodward J.** *Making Things Happen*. New York: Oxford University Press, 2004.
68. **Salmon WC.** *Statistical Explanation and Statistical Relevance*. Pittsburgh, PA: University of Pittsburgh Press, 1971.
69. **Kitcher P.** *Explanatory Unification and the Causal Structure of the World. Minnesota Studies in the Philosophy of Science*. Minneapolis, MN: University of Minnesota Press, 1989, vol. 13.
70. **van Fraassen BC.** *The Scientific Image*. Oxford, UK: Oxford University Press, 1980.
71. **Kaplan DM.** Explanation and description in computational neuroscience. *Synthese* 183: 399, 2011.doi:10.1007/s11229-011-9970-0.
72. **Craver CF**, **Kaplan DM.** Towards a mechanistic philosophy of neuroscience. In: *The Continuum Companion to the Philosophy of Science*, edited by French S, Saatsi J. London: Continuum, 2011, p**.** 268–292.
73. **Salmon WC.** *Causality and Explanation*. New York: Oxford University Press, 1997.
74. **Woodward J.** Explanation**.** In: *The Blackwell Guide to the Philosophy of Science*, edited by Machamer P, Silberstein M. Malden, MA: Blackwell Publishers Ltd**.**, 2002.
75. **Craver CF.** Structures of scientific theories. In: *The Blackwell Guide to the Philosophy of Science*, edited by Machamer P, Silberstein M. Malden, MA: Blackwell Publishers, 2002**.**
76. **Woodward J**, **Hitchcock C.** Explanatory generalizations. I. A counterfactual account. *Nous* 37: 1–24, 2003. doi:10.1111/1468-0068.00426.
77. **Cummins R.** *The World in the Head*. Oxford, UK: Clarendon Press, 2010.
78. **Cummins R.** "How does it work" vs. "What are the laws?" Two conceptions of psychological explanation. In: *Explanation and Cognition*, edited by Keil FC, Wilson RA. Cambridge, MA: MIT Press, 2000, p**.** 117–144.
79. **Cummins R.** Functional analysis. *J Philos* 72: 741–765, 1975. doi:10.2307/2024640.
80. **Piccinini G**, **Craver C.** Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese* 183: 283–311, 2011. doi:10.1007/s11229-011-9898-4.
81. **Bechtel W.** *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York: Taylor & Francis Group, 2008.
83. **Barberis SD.** Functional analyses, mechanistic explanations, and explanatory tradeoffs. *J Cogn Sci* 14: 229–251, 2013. doi:10.17791/jcs.2013.14.3.229.
84. **Cartwright N.** *How the Laws of Physics Lie*. Oxford, UK: Oxford University Press, 1983.
85. **Longino HE.** *Science as Social Knowledge*. Princeton, NJ: Princeton University Press, 1990.
86. **Batterman RW**, **Rice CC.** Minimal model explanations. *Philos Sci* 81: 349–376, 2014. doi:10.1086/676677.
87. **Batterman RW.** *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. New York: Oxford University Press, 2001.
88. **Ross LN.** Dynamical models and explanation in neuroscience. *Philos Sci* 82: 32–54, 2015. doi:10.1086/679038.
89. **Elber-Dorozko L**, **Shagrir O.** Integrating computation into the mechanistic hierarchy in the cognitive and neural sciences. *Synthese* 2019. doi:10.1007/s11229-019-02230-9.
90. **Zednik C.** The nature of dynamical explanation. *Philos Sci* 78: 238–263, 2011. doi:10.1086/659221.
91. **Kaplan DM**, **Bechtel W.** Dynamical models: an alternative or complement to mechanistic explanations? *Top Cogn Sci* 3: 438–444, 2011. doi:10.1111/j.1756-8765.2011.01147.x.
92. **Cao R**, **Yamins DLK.** Explanatory models in neuroscience: part 1—taking mechanistic abstraction seriously (Preprint). *arXiv* 210401489, 2021. https://arxiv.org/abs/2104.01490.
93. **Cao R**, **Yamins DLK.** Explanatory models in neuroscience: part 2—constraint-based intelligibility (Preprint). *arXiv* 210401489, 2021. https://arxiv.org/abs/2104.01489.
94. **Stinson C.** Explanation and connectionist models. In: *The Routledge Handbook of the Computational Mind*, edited by Sprevak M, Colombo M. Abingdon, UK: Routledge, 2018, p**.** 120–133.
95. **Serban M.** The scope and limits of a mechanistic view of computational explanation. *Synthese* 192: 3371–3396, 2015. doi:10.1007/s11229-015-0709-1.
96. **Chirimuuta M.** Explanation in computational neuroscience: causal and non-causal. *Br J Philos Sci* 69: 849–880, 2018. doi:10.1093/bjps/axw034.
97. **Elber-Dorozko L.** Manipulation is key: on why non-mechanistic explanations in the cognitive sciences also describe relations of manipulation and control. *Synthese* 195: 5319–5337, 2018. doi:10.1007/s11229-018-01901-3.
98. **Schrimpf M**, **Kubilius J**, **Hong H**, **Majaj NJ**, **Rajalingham R**, **Issa EB**, **Kar K**, **Bashivan P**, **Prescott-Roy J**, **Schmidt K**, **Yamins DLK**, **DiCarlo JJ.** Brain-Score: which artificial neural network for object recognition is most brain-like? (Preprint). bioRxiv, 2018. doi:10.1101/407007.
99. **Cichy RM**, **Roig G**, **Andonian A**, **Dwivedi K**, **Lahner B**, **Lascelles A**, **Mohsenzadeh Y**, **Ramakrishnan K**, **Oliva A.** The Algonauts Project: a platform for communication between the science of biological and artificial intelligence. *Conference on Computational Cognitive Neuroscience*, 2019. doi:10.32470/CCN.2019.1018-0.
100. **Baumberger C**, **Beisbart C**, **Brun G.** What is understanding? An overview of recent debates in epistemology and philosophy of science. In: *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, edited by Grimm S, Baumberger C, Ammon S. New York: Routledge, 2017, p**.** 1–34.
101. **Stinson C.** From implausible artificial neurons to idealized cognitive models: rebooting philosophy of artificial intelligence. *Philos Sci* 87: 590–611, 2020.
102. **Geirhos R**, **Michaelis C**, **Wichmann FA**, **Rubisch P**, **Bethge M**, **Brendel W.** Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *7th International Conference on Learning Representations (ICLR)*, 2019.
103. **Marr D.** *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W.H. Freeman and Company, 1982.
104. **Chemero A.** *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press, 2009.
105. **Kellert S**, **Longino H**, **Waters CK** (Editors)**.** *Scientific Pluralism. Minnesota Studies in the Philosophy of Science Series*. Minneapolis, MN: University of Minnesota Press, 2006.
106. **Chang H.** *Is Water H_2O? Evidence, Realism and Pluralism*. Berlin: Springer, 2012.
107. **Milkowski M.** *Explaining the Computational Mind*. Cambridge, MA: MIT Press, 2013.
108. **Krakauer JW**, **Ghazanfar AA**, **Gomez-Marin A**, **MacIver MA**, **Poeppel D.** Neuroscience needs behavior: correcting a reductionist bias. *Neuron* 93: 480–490, 2017. doi:10.1016/j.neuron.2016.12.041.
109. **Weiskopf DA.** Models and mechanisms in psychological explanation. *Synthese* 183: 313–338, 2011. doi:10.1007/s11229-011-9958-9.
110. **Tinbergen N.** On aims and methods of ethology. *Zeitschrift für Tierpsychologie* 20: 410–433, 2010. doi:10.1111/j.1439-0310.1963.tb01161.x.
111. **Tsao DY**, **Livingstone MS.** Mechanisms of face perception. *Annu Rev Neurosci* 31: 411–437, 2008. doi:10.1146/annurev.neuro.30.051606.094238.
112. **Chang L**, **Egger B**, **Vetter T**, **Tsao DY.** Explaining face representation in the primate brain using different computational models. *Curr Biol* 31: 2785–2795, 2021. doi:10.1016/j.cub.2021.04.014.
113. **Leavitt ML**, **Morcos AS.** Selectivity considered harmful: evaluating the causal impact of class selectivity in DNNs. *International Conference for Learning Representations (ICLR)*, 2021.
114. **Kendi IX.** *How To Be An Anti-Racist*. New York: Penguin Random House, 2019.

115. **Doran D**, **Schulz S**, **Besold TR.** What does explainable AI really mean? A new conceptualization of perspectives (Preprint). *arXiv* 171000794, 2017. https://arxiv.org/abs/1710.00794.
116. **Lipton ZC.** The mythos of model interpretability. *ICML 2016 Workshop on Human Interpretability in Machine Learning (WHI)*, 2016.
117. **Haufe S**, **Meinecke F**, **Görgen K**, **Dähne S**, **Haynes J-D**, **Blankertz B**, **Bießmann F.** On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage* 87: 96–110, 2014. doi:10.1016/j.neuroimage.2013.10.067.
118. **Buckner CJ.** Black boxes, or unflattering mirrors? Comparative bias in the science of machine behavior. *Br J Philos Sci*. In press. doi:10.1086/714960.
119. **Baraniuk R**, **Donoho D**, **Gavish M.** The science of deep learning. *Proc Natl Acad Sci USA* 117: 30029–30032, 2020. doi:10.1073/pnas.2020596117.
120. **Donoho D**, **Raghu M**, **Rahimi A**, **Recht B**, **Gavish M.** The Science of Deep Learning (Online). 2019. http://www.nasonline.org/programs/nas-colloquia/completed_colloquia/science-of-deep-learning.html [2020 April 30].
121. **Metz C.** Google Researchers Are Learning How Machines Learn (Online). *New York Times*, March 6, 2018. https://www.nytimes.com/2018/03/06/technology/google-artificial-intelligence.html.
122. **Morcos AS**, **Barrett DGT**, **Rabinowitz NC**, **Botvinick M.** On the importance of single directions for generalization (Preprint). *arXiv* 1803.06959, 2018. https://arxiv.org/abs/1803.06959
123. **Olah C**, **Satyanarayan A**, **Johnson I**, **Carter S**, **Schubert L**, **Ye K.** The Building Blocks of Interpretability. Distill 2018. doi:10.23915/distill.00010.
124. **Sundararajan M**, **Taly A**, **Yan Q.** Axiomatic attribution for deep networks (Preprint). *arXiv* 1703.01365, 2017. https://arxiv.org/abs/1703.01365.
125. **Kornblith S**, **Norouzi M**, **Lee H**, **Hinton G.** Similarity of neural network representations revisited (Preprint). *arXiv* 1905.00414, 2019. https://arxiv.org/abs/1905.00414.
126. **Sussillo D**, **Barak O.** Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Comput* 25: 626–649, 2013. doi:10.1162/NECO_a_00409.
127. **Leavitt ML**, **Morcos AS.** Towards falsifiable interpretability research (Preprint). *arXiv* 2010.12016, 2020. https://arxiv.org/abs/2010.12016.
128. **Nelson LH.** Feminist philosophy of science**.** In: *Blackwell Guide to the Philosophy of Science*, edited by Machamer P, Silberstein M. Malden, MA: Blackwell Publishers, 2002, p**.** 312–331.
129. **Nelson LH.** *Who Knows: From Quine to a Feminist Empiricism*. Philadelphia, PA: Temple University Press, 1990.
130. **Harding S.** Rethinking standpoint epistemology: what is "strong objectivity?" *Centennial Rev* 36: 437–470, 1992. http://www.jstor.org/stable/23739232.
131. **Devezer B**, **Nardin LG**, **Baumgaertner B**, **Buzbas EO.** Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLoS One* 14: e021612, 2019. doi:10.1371/journal.pone.0216125.