Trends in Cognitive Sciences

CellPress
REVIEWS

Opinion

Deep Neural Networks as Scientific Models

Radoslaw M. Cichy[1,2,3,*] and Daniel Kaiser[1]

Artificial deep neural networks (DNNs) initially inspired by the brain enable computers to solve cognitive tasks at which humans excel. In the absence of explanations for such cognitive phenomena, in turn cognitive scientists have started using DNNs as models to investigate biological cognition and its neural basis, creating heated debate. Here, we reflect on the case from the perspective of philosophy of science. After putting DNNs as scientific models into context, we discuss how DNNs can fruitfully contribute to cognitive science. We claim that beyond their power to provide predictions and explanations of cognitive phenomena, DNNs have the potential to contribute to an often overlooked but ubiquitous and fundamental use of scientific models: exploration.

Highlights

Neurally inspired deep neural networks (DNNs) have recently emerged as powerful computer algorithms tackling real-world tasks on which humans excel, such as object recognition, speech processing, and cognitive planning.

In the absence of scientific explanations regarding how humans solve such tasks, some cognitive scientists have turned to DNNs as models of human brain responses and behaviour.

In visual and auditory processing, DNNs were found to predict human brain responses and behaviour better than other models.

The use of DNNs as models in cognitive science has created a heated debate about their scientific value: in particular, are DNNs only valuable as predictive tools or do they also offer useful explanations of the phenomena investigated?

The Contested Value of Deep Neural Networks in Cognitive Science

In recent years, neurally inspired [1,2] artificial **deep neural networks** (DNNs; see Glossary) have revolutionised first computer vision [3] and subsequently other domains such as natural language processing [4], control and planning (such as playing games, e.g., Atari and Go [5,6]), and navigational tasks (such as finding the shortest path on a subway map [7]).

DNNs are computational models consisting of many simple processing units (akin to neurons) that work in parallel and are arranged in interconnected layers. Simple neural networks consist of an input layer and an output layer; when more layers are stacked, the networks are called deep [8,9]. A DNN learns to perform particular tasks through training, during which the strength of connections between units is learned. Subsequently, the trained DNN is used to perform the same task on novel inputs.

While DNNs were built with an engineering goal in mind, for example, to make a computer solve a particular classification problem, recruiting DNNs for cognitive science might yield useful insights. In this spirit, researchers have shown that DNNs predict human perceptual similarity judgements [10,11] and outperform any other model in accounting for neural activity in primate sensory cortices [12–18] (Figure 1).

However, this use of DNNs has sparked heated debate over its scientific value (Box 1). The goal of this opinion article is to evaluate what is at stake from a bird's-eye view. Here, we discuss how philosophy of science [19] informs the debate on DNNs (Figure 2). Specifically, we first put DNN models into a broader context, arguing that they find their place in the plurality and diversity of models in cognitive science. Next, we discuss DNNs in the context of the two main goals of science: prediction and explanation [20]. We claim that DNNs have diverse virtues for both goals. Finally, we draw attention to exploration [21] as a fundamental modelling practice. We claim that by virtue of their excellent exploratory potential DNNs can play an important role in scientific progress.

The Nature of Good Scientific Models

To evaluate whether DNNs are good scientific models, we need to agree on what makes a good model in the first place. Can we formulate a list of properties good models fulfil and check DNNs

[1]Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany
[2]Berlin School of Mind and Brain, Humboldt-Universität Berlin, Berlin, Germany
[3]Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany

*Correspondence: rmcichy@gmail.com (R.M. Cichy).

Box 1. Arguments Pro and Contra DNNs in Cognitive Science

Arguments of critics and proponents can be ordered into five points (for an excellent historical exposition, see chapter 4 in [1]).

Overall potential: Critics are pessimistic about the outlook of DNNs in cognitive science. They draw attention to their limitations, and the divergence between the limited range of tasks a DNN can do today, and the many cognitive functions lacking explanation [45,80]. Fundamentally different approaches might be needed to solve many cognitive tasks. Proponents stress the potential unprecedented opportunities [49,63]. They see promise for a new convergence of artificial intelligence (AI) and neuroscience/psychology, and a new framework for relating brain function and complex real-world behaviours [48,79,81,82].

Explanation: Conceding that DNNs might predict brain activity or behaviour well, contrasting prediction and explanation, critics argue that DNNs fall short of explaining the phenomenon [45]. Proponents argue that DNNs do explain either in the same way as other models, in a qualitatively different way [13,15,36,49], or that DNNs afford rich post hoc explanations.

Interpretation: Related to explanation, critics characterise DNNs as 'black boxes', where the contribution of individual model components remains untraceable [45,80]. Thus, using DNNs to explain biological brains means exchanging one opaque system for another, without additional gain. Proponents, while conceding that DNNs are not immediately transparent, argue that this does not ultimately preclude their interpretation [83]. DNNs offer better access to experimentation than humans as they can be investigated swiftly and cheaply using a growing set of techniques to make them transparent (often termed *in silico* electrophysiology) [62].

Biological realism: While conceding that DNNs are inspired by biological neural networks, critics argue that DNNs lack biological realism. For example, the DNN units mimicking neurons are highly abstracted and lack most of the dynamics of biological neurons. Proponents instead point out that abstraction and idealisation are essential to any modelling endeavor and thus no *a priori* reason to reject a model [49,63]. Furthermore, future developments may increase the functional similarity between DNNs and the biological brain.

Scientific validity: Critics say that the current use of DNNs to learn about biological systems is unscientific because it is untheoretical. DNNs were not made to model brains or human behaviour, and they are not deducted from theories to test hypotheses. Proponents argue that the origin of models is irrelevant to their value in science [84]. Instead, they should be judged by their power for explanation or prediction only.

against these properties? Or, can we select a particular model that we believe to be an excellent scientific model as a standard and compare DNNs against it? Reflecting on the use of models in a broader scientific context, we argue that the case is not as simple as that.

No One Perfect Model: Necessary Trade-Offs

One way to determine what makes a good model is to define what would make the best possible model and then compare. Can we have that one, perfect model for any one phenomenon of interest in cognitive science? Is the current state of model diversity therefore an interim stage to be overcome? We claim that this is unlikely, as favoured properties of models in cognitive science (i.e., **desiderata**) have to be traded against each other [22], and no single model can fulfil all of them.

Desiderata fall into two classes: theoretical and non-theoretical. Theoretical desiderata for models of biological phenomena are precision (how exact is the model outcome), realism (how similar is the model to the phenomenon), and generality (how well does it generalise from one case to another) [23]. If the class of phenomena to be modelled is fully homogenous, that is, all members of the class have the same properties, precision, realism, and generality can be achieved simultaneously. However, when this is not the case, a model cannot be fully precise (model each member of the class correctly) and general (model all members) (http://philsci-archive.pitt.edu/1000/; [24,25]).

Glossary

Analogy: a similarity between relations in two different domains.
Box-and-arrow model: a model of information processing in which boxes represent components of an information processing system, and arrows represent information flow between those components.
Deep neural network: a computer algorithm inspired by biological neural networks, consisting of units akin to neurons and defined in function by the connection between the units (Figure 1A). Units are often not connected to all other units but ordered in layers. Layers between the input and output layers are called hidden layers, and neural networks with hidden layers are called deep.
Desiderata: plural of desideratum (lat.), denoting something that is desired as essential.
Explanandum: a description of the phenomenon to be explained.
Explanans: a set of sentences cited as evidence for the phenomenon to be explained, including a natural law or statistical regularity.
Inferior temporal (IT) cortex: the inferior part of temporal cortex in non-human primates and part of the ventral visual stream. It is the end stage of a sequence of cortical processing steps starting in visual area V1. It is most prominently described as harbouring cells that respond relatively specific to particular objects or object categories but tolerant to changes in viewing conditions such as size, angle, and lighting. The putative functional analogue in humans is the lateral occipital complex, a set of regions responding more strongly to images of intact objects than to scrambled objects.
Ready-made: an art term (strongly connected to Marcel Duchamp) designating an everyday object that was made for some purpose, found by an artist, and chosen as a piece of art with little or no modification.
Teleological: from Greek *telos* (end, goal, purpose), related to a goal, aim, or purpose.
Two-stream hypothesis: hypothesis put forward on the basis of neuropsychological data. It divides the visual brain into a ventral and dorsal stream. The ventral stream is meant to process the content of

visual scenes ('what' is perceived), whereas the dorsal stream mediates knowledge of 'where' something is perceived, or 'how' it can be used.

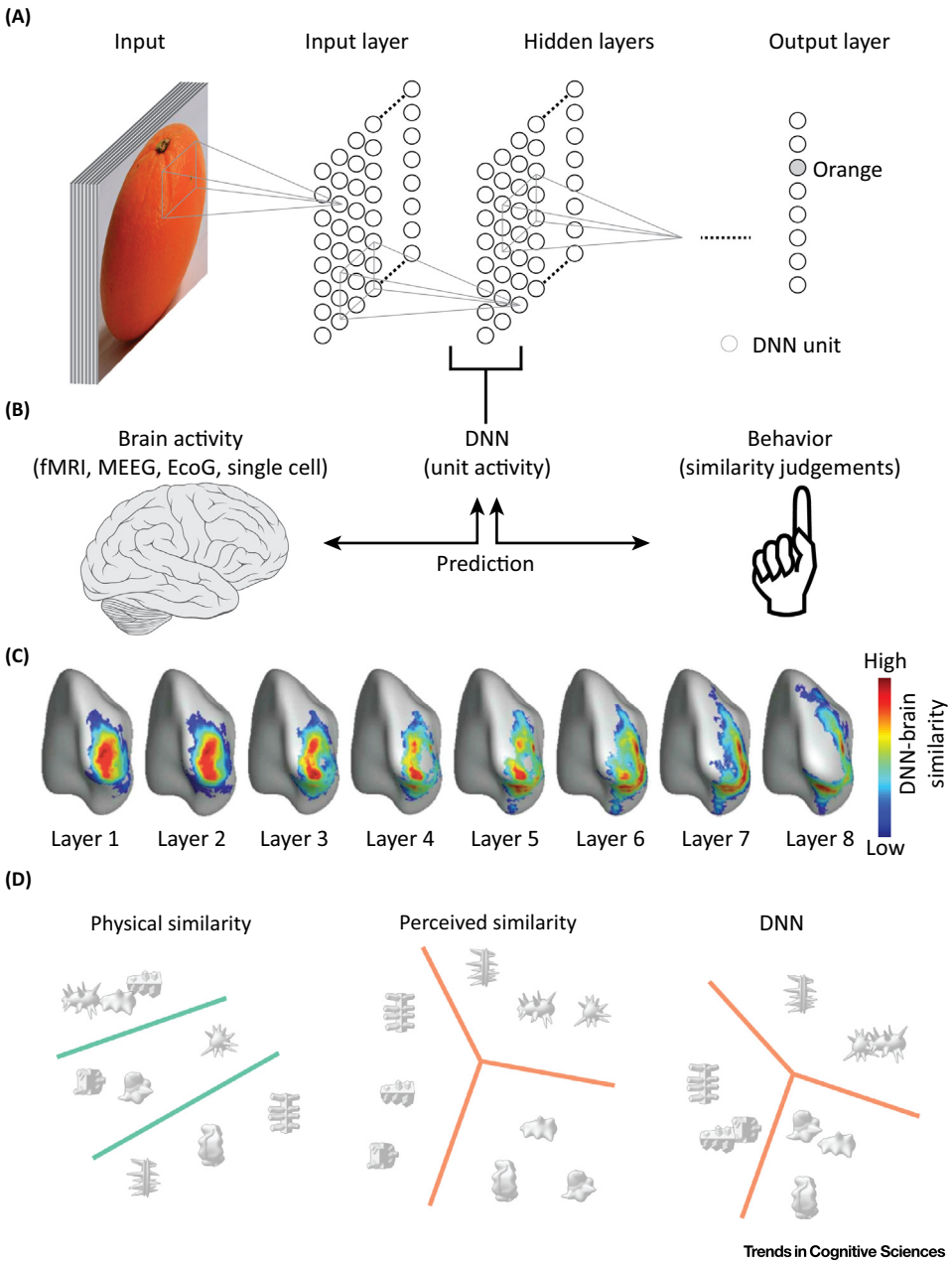

Figure 1. Rationale and Example Uses of DNNs in Cognitive Neuroscience. (A) Example of a deep neural network (DNN) used to categorise objects. DNNs are computer algorithms inspired by biological networks. They consist of artificial units akin to neurons. Units are connected to each other, and the connection weights determine what a unit computes, that is, what input features it represents. Units are usually ordered in layers. An input layer is activated directly by the input (here, pixel values), and an output layer produces a response given the task the DNN was trained on (here, a category label). Networks that have layers between input and output layers, called hidden layers, are called deep. Units between layers can be connected feed-forward and feed-back, and within a layer in recurrent manner. During training, the DNN learns features needed to carry out the task, rather than the modeller setting the features explicitly (for introductory reviews, see [36,49,63,79]). (B) Researchers in cognitive science use DNNs to predict brain activity and behaviour. (C) Example DNN-brain activity relation. An eight-layer DNN trained on object categorisation was compared to fMRI brain activity. There was a hierarchical relationship: early layers of the DNN predict low-level visual brain regions, and later layers

(Figure legend continued on the bottom of the next page.)

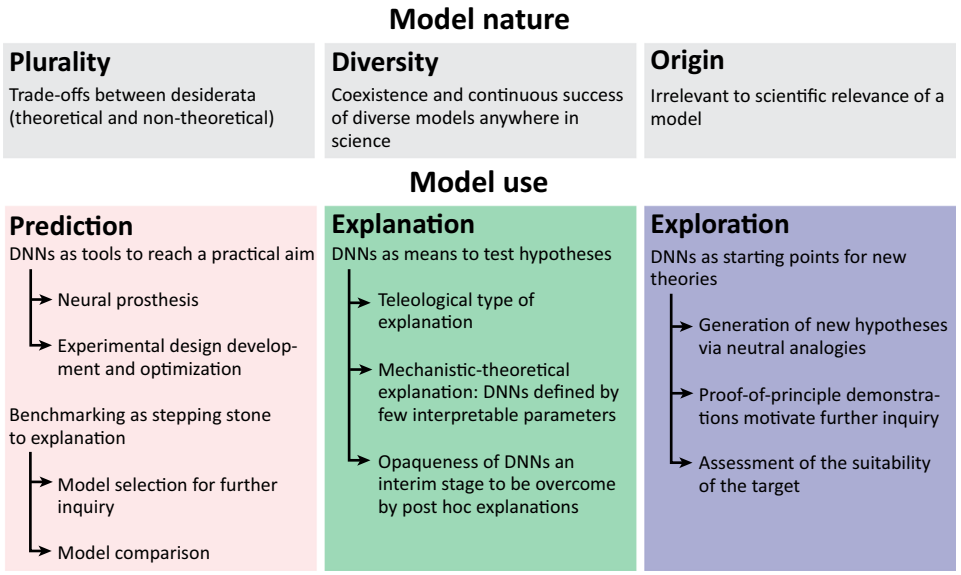

Figure 2. Deep Neural Networks in the Light of Considerations about Model Nature and Model Use in Science. Assessing model nature, we argue that there are many models of the same phenomenon (plurality), that these models likely differ from each other (diversity), and that where the model comes from is irrelevant for its scientific value (origin). DNNs fit into this as inhabiting one particular niche in the world of models in cognitive science. Assessing model use, we argue that DNNs hold promise for all three functions of models in science: for prediction, for explanation, and for exploration.

The behaviour and brains of biological beings are not homogeneous at any scale: they are different between species and individuals, and they change with experience and across the lifespan. Thus, there cannot be one model that is perfectly precise and general, or realistic and general simultaneously. Instead, there will be many models fulfilling desiderata differentially well. For example, we can either model vision for a particular person, or across a group of people, or for a diverse set of organisms (e.g., humans and spiders), but not for all of them at once. Making the model's target and the theoretical desiderata explicit is thus necessary to enable proper judgement about modelling success and its comparison to other models. Transferred to the debate on DNNs, critique or praise about a DNN's achievement can only be appreciated with respect to the desiderata pursued and what was intended to be modelled with what aim.

In scientific practice, non-theoretical desiderata, that is, practical considerations such as speed of computation, ease of manipulation, and ethical considerations, often take precedence over theoretical desiderata. Imagine we had a model of the brain that fulfils all theoretical desiderata but is too slow to compute interesting results at relevant time scales. In this case, a model that sacrifices theoretical desiderata for speed is scientifically preferred. As another example, consider animal models: while for theoretical reasons (e.g., realism) an animal might be a suitable model for humans, ethical limitations on animal experiments can pose severe

predict higher level regions [13]. (D) Example DNN-behaviour relation. A DNN trained on object categorisation predicted the perceived (similarity judgments) rather than the physical similarity (i.e., pixel values) of visual stimuli [10]. Abbreviations: EcoG, electrocorticography; MEEG, magnetoencephalography/electroencephalography.

constraints. DNNs excel in these non-theoretical desiderata: they compute cheaply and swiftly, and their investigation has fewer ethical limits than many other models.

DNNs as One of a Diverse Set of Models

Given the need for many different optimal models (relative to a set of desiderata), will all those models ultimately be of the same kind, for example, all DNNs, or will they be diverse? We offer a plausibility argument. Bewilderingly diverse sets of models coexist and enjoy continued success across scientific disciplines that are at different degrees of maturity. This set includes single equations (e.g., the Ising model [26] in quantum physics), physical scale models that are copies of the phenomenon (e.g., miniature bridges in engineering), and complex computational models (e.g., climate or economic models). In cognitive science, the set ranges from **box-and-arrow models** (e.g., Baddeley's model of working memory [27]) to computational models (e.g., symbolic [28], neural networks [1], dynamical systems [29]) to whole animals (e.g., Aplysia or mice). The coexistence and continued success of diverse models across scientific disciplines and at different degrees of maturity speaks against the idea of a single best kind of model. Instead, it argues for embracing the plurality and diversity of models [30,31].

Transferred to the debate on DNNs, this suggests that at any time a single unified model unlikely will do all the work in cognitive science but that a multitude of different models will. We thus plea to neither dismiss DNNs prematurely nor to envision them as the kind of model that cognitive science will ultimately progress to.

The Origin of Models Is Scientifically Irrelevant

In spite of their plurality and diversity, models are often considered scientific by virtue of a common origin: they are all derived from a theory to instantiate or test that theory. However, this proposal squares badly with scientific practise. Models are rarely straight-forwardly deduced from theory: one cannot push a button in the theory and a model pops out ('vending machine view of models') [32]. More similar to the artistic process than logical deduction, there is no codified set of rules in model building. Instead, the process often involves creativity, chance, and transfer [24]. Furthermore, as mentioned above, model building involves fulfilling non-theoretical desiderata such as simplification and approximation.

The DNNs used today in cognitive science are not derived from theory to test particular predictions of that theory. Admittedly, DNNs are loosely inspired by classical theories of biological brain function, but they were not built to test those theories. Therefore, the hijacking of DNNs by cognitive scientists is more analogous to the use of **ready-mades** in art. However, this is no reason to dismiss the model, as the origin of DNNs plays no role for their scientific value.

In sum, putting DNNs as models in cognitive science into the broader context of scientific modelling in general, we made three observations: (i) all models of cognitive phenomena constitute trade-offs between desiderata, and also DNNs offer unique strengths and weaknesses; (ii) the set of models used in cognitive science is diverse, and DNNs are one particular, useful kind; and (iii) the origin of a model is irrelevant for its scientific value, and so are the DNNs' origins. Having established their place among a diverse set of models, we discuss how DNNs fare in scientific prediction, explanation, and exploration.

The Predictive Power of DNNs

In technology and engineering the primary goal is to create artefacts that do things (i.e., correctly predict a particular outcome), while explanation often takes second place [33]. Are

DNNs useful in cognitive science in this way? While DNN critics concede the predictive power of DNNs, they often dismiss it as less valuable for science than explanation. Here, we affirm the value of prediction on two grounds.

A pragmatic reason is that due to its predictive power a DNN could be used akin to a tool to reach practical aims without direct recurrence to explanation. One promising future direction is medical applications. Imagine a patient that has a lesion in visual cortex, for example, through stroke, and consequentially suffers from a loss of object vision (visual agnosia). For this patient, an artefact predicting brain activity in the damaged cortex, such as a DNN, could serve as a prosthesis. The scientific benefit is in predicting and thus substituting brain function; explanation and understanding are secondary. Admittedly, current DNNs are far from perfect in mimicking the full neural dynamics of visual cortex and object recognition behaviour [34] that would be necessary to restore visual functions faultlessly. They roughly model neural signals about an object's category, location, or pose [35]. However, for patients facing a sudden loss of object vision, even an imperfect restoration of such basic abilities would be medically highly relevant.

Another promising application is the non-invasive experimental control of brain activity [36]. Recent research has demonstrated the feasibility of using DNNs for systematically manipulating brain activity in visual regions V1 [37] and V4 [38]. By using DNNs to predict visual brain responses, the authors were able to synthesise images that drove neural ensembles into predetermined desired states. This experimental control over brain activations promises new insights into neuronal functioning. For example, consider the interplay of multiple brain regions during sensory processing. Using optimal stimuli derived from DNN analyses allows for tightly controlling activity in one region while establishing the workings of another region.

A second, more theoretical reason is that predictive power can be a stepping stone towards explanation. Consider the benchmarking of different models with respect to how well they predict brain activity [12,15,39]. For one, this comparison allows cognitive scientists to preselect models performing well as promising candidates for further inquiry [15]. Second, comparison of models differing in success might reveal factors relevant for their success and thus contribute to explanation [12,13,15,39,40]. Third, predictive power ultimately is a criterion for successful explanation: we cannot be satisfied with a model that offers an explanation but also does not predict well [41,42]. Current DNNs thus constitute a clear, quantitative challenge to current and future models that aim at offering scientific explanations.

In sum, we argue that the predictive power of DNNs is useful for cognitive science in two ways: it allows DNNs to be used to pursue practical aims such as experimental control or neural substitution, and it is a stepping stone towards explanation. We turn next to how DNNs can do explanatory work proper.

The Explanatory Power of DNNs

How can a model do explanatory work in cognitive science (Box 2)? The blueprint notion many researchers have in mind is so-called mathematical–theoretical modelling [43,44]. There, a few relevant variables for describing a phenomenon are identified, it is hypothesised how they interact, and the variables and their interaction are modelled mathematically. Each variable is *a priori* linked to a part of the phenomenon modelled in the world, such that changes in the model variable are directly interpretable: the model is transparent.

Compared to this class of models DNNs indeed look different. A DNN often has millions of parameters that are learned by the network rather than being set *a priori*. It is therefore not

Box 2. The Logical Form and Concept of Explanation

In everyday practise, scientists conduct experiments and interpret the results such that other scientists acknowledge them as good explanations. However, what makes an argument a good explanation is not easy to describe. There is a diversity of notions differing across disciplines and contexts [85,86]. How do DNNs fit in?

The deductive-nomological notion [87]. An explanation consists of an **explanans** and an **explanandum**. The explanandum must follow logically from the explanans (the deductive part). The explanans must be true, and it must contain a law of nature necessarily (the nomological part). However, few scientists, and cognitive scientists are no exception, formulate their arguments explicitly in this form, and it is unclear whether or how to apply laws of nature in the context of evolved biological beings and their cognition.

The inductive-statistical notion [88]. The explanandum follows probabilistically rather than logically from the explanans (the inductive part), and the explanans contains a statistical regularity (the statistical part). While cognitive scientists working with DNN do not explicitly use this form, it is conceivable that their statements could be transformed into such a form.

The causal mechanistic notion [89]. The rationale is that to explain a phenomenon is to give its cause. The explanation enumerates the causal processes and how they interact up and lead to the phenomenon. Cognitive scientists working with DNNs are beginning to use this notion in the context of experimental interventions rather than observations [38].

The unificationist notion [90,91]. Explanation amounts to unify different phenomena in a common account, showing connections and relationships between phenomena whose relation was previously unclear. Ideas of DNNs as general frameworks for modelling brain function and cognition [48,49,63] are reminiscent of this notion.

The pragmatic notion [43,92]. Pragmatists stress that whether an explanation is successful, depend irreducibly on facts about the interests, goals, and beliefs of those providing or receiving explanation. This is directly relevant to DNNs in cognitive science given the interdisciplinarity of the endeavour. Different disciplines (such as psychology, biology, and computer science) have different standards and criteria of success. Thus, we need to acknowledge and deal with such differences to fairly assess expounded arguments.

immediately apparent how the variables map onto the world and how they interact among each other. Without fully understanding its inner workings, the DNN appears as a black box. This poses a challenge for explanations based on DNNs: one cannot explain one black box by another [45].

To make the distinction more concrete, consider research on visual representations in primates. Mathematical–theoretical modelling identified a few variables, for example, spatial filters or geometrical primitives [46,47] that can be directly linked to cortical activations. By contrast, DNN-based modelling yields no *a priori* mapping between model parameters and neurons; a DNN's many parameters are learned during training, in a non-transparent way. From the perspective of the mathematical–theoretical modeller, using DNNs to predict brain activations (Figure 1) may appear questionable.

Should we therefore conclude that DNNs have no explanatory power? There are at least three mutually non-exclusive reasons why we should not.

The first argument is that DNNs do provide an explanation but of a qualitatively different kind [15,48]. The answer to a question such as 'why does a DNN unit behave such and such' is not 'because it represents this or that feature of the world', but 'because the unit needs to respond such in order to fulfil its function in enabling a particular objective, such as object recognition'. That is, the nature of the explanation is **teleological**.

A second perspective is that appearance is deceptive, and DNNs are effectively used in the same way as traditional mathematical–theoretical models. After all, DNNs are fundamentally

defined by a handful of parameters set *a priori*, such as architecture, training material, training procedure, and objective [13,49]. These variables directly refer to specific phenomena in the world, the model is thus transparent and has explanatory power. The fact that the handful of variables in DNNs is different from the handful of variables in other models is not a problem by itself but follows from the fact that models differ in what aspects of the phenomenon they emphasise, and which they hide (Box 3) [50]. Furthermore, the fact that historically DNNs were first compared to the brain without explicit identification of the crucial parameters relevant for explanation is neither a principled reason that DNNs cannot provide explanations of this kind. Last, the fact that a parameter such as 'training material' is in fact a very complex set of data does not defeat its use as a single, simple parameter in an explanation. In the context of DNN modelling, it reflects one single relevant part of the DNN modelling procedure. Such an abstraction from fine-grained levels of description into single parameters at higher levels of description is common practice in the modelling of complex phenomena. Naturally, this in no way precludes the possibility of explanatory attempts at a finer grained level of description.

A third perspective is that immediate opaqueness is only an intermediary stage that will be overcome by DNN models' strong potential for post hoc explanations. There is a growing method repertoire to make DNN computations explicit through visualisation, text description, or finding representative examples [51–61]. Uncovering the mapping between model variables and the world could make DNNs transparent and thus explanatory. As such, DNNs in cognitive science appear no different than model organisms in biology [62]. While model organisms are opaque and the transfer of insight from model to target phenomenon requires additional work, their explanatory value in providing experimental access is well acknowledged. Note also that complex (and thus immediately opaque) models might be inevitable for modelling complex

Box 3. Models Are Not Neutral Tools

Models are not neutral tools that we use at will, and DNNs are no exception. Instead, DNNs shape the way we think and do science.

Models emphasise some aspects of the world, while hiding others. Models that are computationally well-developed formalisms allow us to do particular things easily, while others are hard to achieve. For example, consider programming languages: what is easily expressed in one formalism can be cumbersome in another. Analogously, two models of the same phenomenon, such as a DNN and a mathematical–theoretical model, may enable and restrain different ways of thinking, respectively [1,24,50].

Models shape the perception of problem situations and what counts as a solution. A model's particular affordances and constraints affect how users perceive problem situations, on what they focus attention, and what they perceive as a solution [21]. In the way that for a man with a hammer everything is a nail, for a scientist with a DNN every problem may be solved by the DNN, for example, by learning an implicit solution from large data rather than through explicit engineering. Consider for example the contingency and historicity of the proposed minimal criteria for future modelling of sensory cortex, formulated after DNNs were shown to outperform other models in that domain [36]: models are to be stimulus-computable (model accepts arbitrary stimuli from target domain), mappable (model components map onto parts of the neural system), and predictable (models provide stimulus-specific predictions for neural activity). These criteria happen to be the criteria that current DNNs fulfil. They are reasonable in so far as that future research shall not stay behind the state of the art. However, their force derives from the availability of a model that fulfils them today, not from theory. If hypothetically a DNN had been found that excelled on another dimension, for example, biological realism using spiking units, that dimension would have reasonably been claimed a minimal criterion.

Models shape social reality. Once technological artefacts are commonly used, they are not mere tools to realise pre-defined scientific goals but begin to shape social reality in a way that affects the user's desires and interests. Scientific models such as DNNs are no exception. The demand for skilled DNN researchers influences graduate programs. The rise of DNNs in one field (e.g., computer vision) suggests analogous solutions in other fields (e.g., AI and cognitive science) [8], creating further demand for education, and for novel venues to discuss research [93]. Thus, DNNs shape what companies want to invest in, students want to learn, and researchers want to research.

phenomena such as human behaviour and brain activity that depend on large amounts of world knowledge that cannot be easily summarised [63]. Simpler (and thus immediately more transparent) models might not be able to capture that knowledge. The goal must therefore be to reduce opaqueness, but not at the price of predictive power.

Together, we presented three different perspectives from which DNNs have explanatory power: (i) they provide teleological explanations; (ii) despite their deceptive appearance, they provide the same explanations as traditional mathematical theoretical models; and (iii) owing to their complexity, they have strong potential for post hoc explanations. We now turn to a third, often underappreciated role of scientific models beyond prediction and explanation: exploration.

The Exploratory Power of DNNs

An idealised view of natural science is that it proceeds by deriving hypotheses from a theory and testing them in experiments. But what to do if a fully-fledged and convincing theory is missing? Then we need to explore [64,65] to create starting points for new theories, rather than predict or explain. This means a shift of perspective from models as tools for predictions, or akin to theories for explanation, to exploration [21].

Observing the scientific practise indicates that exploration is an omnipresent strategy. We learn from models by building and manipulating them. Modellers play with their models, exploring how they behave and getting a 'feel' for them [66].

How do DNNs in cognitive science fit into this picture? We claim that they contribute to exploration exceedingly well. First, the absence of a fully-fledged theory of any single cognitive function gives models an important exploratory role. Second, their complexity makes DNNs particularly well suited for exploration: there is lots to explore.

We highlight three ways how exploration using DNNs benefits cognitive science: (i) by generating new hypotheses that can be investigated empirically, (ii) through proof-of-principle demonstrations that create plausibility, and (iii) by determining the suitability of the target phenomenon.

DNN Exploration Generates New Hypotheses

Exploring models can create novel ideas. How does this formally work? Mary Hesse offered a canonical analysis in terms of **analogies** [67]. Any model has positive, negative, and neutral analogies with respect to the target modelled. Positive and negative analogies are characteristics that are known to be shared or not shared, respectively, by the model and the target. Neutral analogies are those for which we do not know whether the model and the target share them. Negative analogies are due to the distortions, idealisations, and abstractions that limit a model's scientific value. While positive and negative analogies are established relations between model and target, and thus offer no new insight, neutral analogies allow us to learn novel facts about the target.

Transferred to the debate on DNNs, this means that the investigation of neutral analogies between DNNs and the brain is a promising source of new hypotheses for empirical testing. For example, like the brain, DNNs consist of rather simple processing units whose concerted activation enables complex functions such as object recognition (positive analogies). However, while DNNs are made in silica and their units have simple time-invariant dynamics, human brains are organic and show highly complex dynamics (negative analogies). The neutral analogies comprise all properties of which we do not know whether DNNs and brains share them, or not.

DNNs as Proof-of-Principle Demonstrations

In engineering, proof-of-principle demonstrations show the feasibility of a particular approach to a problem by creating an artefact that solves the problem. Their success motivates further scientific attention.

Transferred to the debate on DNNs, proof-of-principle demonstrations highlight the feasibility of modelling behaviour or neural data using DNNs and thereby motivate further scientific attention towards DNNs.

To exemplify, feed-forward DNNs trained on object categorisation reach performance levels similar to humans on particular object recognition tasks, and they accurately predict object-related brain activity [12,13,15]. These results are proof-of-principle demonstrations for the feasibility of purely feed-forward, bottom-up rather than feed-back, top-down dependent solutions. While they do not definitely prove that object vision solely relies on bottom-up processing, such investigations motivate further research exploring how much of vision a purely bottom-up approach can account for. Conversely, if particular analogies between DNNs and the brain only emerge when top-down wiring is implemented in the DNN model [68–70], one could infer that top-down processes are also recruited during neural processing.

DNNs Assess the Suitability of the Phenomenon Modelled

In theories that are not fully fledged, the concepts that identify and distinguish parts of the phenomena to be explained are not fully established either. In this case, experimentation and theoretical concept development are intertwined activities [71], where concepts (operationalisations) are refined and revised in the light of experimental results.

Modelling can thereby have the same effect as experimentation [72,73]: it can make us change our concepts. For example, by observing the model's behaviour, we find novel regularities that may become part of the concept. Or, by determining under which conditions modelling results are reproducible, we engage in concept stabilisation [64]. In effect, the concepts we use to identify and describe a phenomenon change.

To transfer this to the debate on DNNs, consider two examples in the investigation of the neural basis of object vision. In the first example [35], the authors explored how a DNN trained on object categorisation predicts different object properties. As expected, object category was increasingly well predicted along the DNN's processing hierarchy. Given that successful object recognition requires tolerance against changes in category-orthogonal properties (such as location or size), it seemed plausible to assume that the prediction of such orthogonal object properties diminishes along the processing hierarchy. Surprisingly, the authors found the opposite: the DNN predicted category-orthogonal object properties increasingly well along the DNN's processing hierarchy, too. This led to a hypothesis challenging current neural theories of object vision: **inferior temporal (IT) cortex** in primates (thought to represent object category across viewing conditions) may be representing such category-orthogonal properties, too. Analysis of electrophysiological data confirmed this prediction. Here, model exploration and subsequent empirical investigation led to an important refinement of the classical **two-stream hypothesis** of the visual brain [35].

Another example involves recent findings on transfer learning, where a DNN trained on one task (usually object categorisation) is evaluated in another task [74–78]. During transfer learning, typically the pre-trained DNN's output layer is replaced; next, the mapping between DNN activations and the output layer is re-trained on another task. This procedure tests whether features learned during

one task are general enough to solve another task. It turns out that a DNN trained on object categorisation learns features useful for a wide range of tasks, such as fine-grained discrimination, saliency prediction and scene recognition. Such observations make predictions to which degree resources underlying different visual functions may be shared in the human brain.

Caveats and Limitations of DNN Exploration

The explorative use of models is subject to several caveats and limitations. First, when modelling precedes theory, quality standards and benchmarks are not well developed and often implicit. We plea to give DNNs in cognitive science the benefit of the doubt, avoiding too strict standards that may curb bourgeoning developments prematurely. In turn, DNN modelling has to use this leeway for theory building. Second, the same model may be used exploratively in one context (e.g., to generate new hypotheses), but for explanation in another (e.g., as an explicit model of neuronal functioning). To enable fair assessment, researchers should make transparent how they use the model [30]. Third, exploring models rather than the world bears the danger of mistaking the model for the world. Investigating DNNs might be mistaken for investigation of the human brain or behaviour. To avoid this, exploration must be accompanied by experimentation on the target phenomenon.

In sum, the exploratory power of DNNs makes them excellent starting points for new theories and for revising existing theories. DNN modelling generates new hypotheses, it allows for proof-of-principle demonstrations, and it helps refine our scientific concepts. To not lead astray, however, exploration should not be confused with explanation or experimentation.

Concluding Remarks and Future Directions

Taking a bird's-eye view from the stance of philosophy of science, we took a fresh look on what is at stake in the debate on DNNs as models for behaviour and neural activity. We emphasise four take-home messages for future research (Figure 2). First, given the current level of theory development and the need to trade-off model desiderata, we should embrace DNNs as one of many diverse kinds of useful models. Second, through their predictive power DNNs have rich potential as tools for scientific research and application. Third, we should use DNNs' explanatory power for theorisation, but make explicit what type of explanation is at stake to allow fair assessment and criticism. Finally, the exploratory power of DNNs deserves our heightened attention. The computational complexity of DNNs makes them particularly suited candidates for exploration, promising theoretical insight that we cannot foresee today.

The development of DNNs is progressing rapidly, and the applications of DNNs in cognitive science are increasing and diversifying. This means that the types of predictions, explanations, and explorations employed will diversify accordingly (see Outstanding Questions). Theoretical discussion accompanying these developments is needed to ensure that cognitive scientists make the best possible use of DNNs models.

Outstanding Questions

We outlined that particularly neutral analogies (unknown correspondences between model and phenomenon) and negative analogies (known divergences between model and phenomenon) drive modelling development. For DNNs, which of these analogies are relevant for improving the fit between DNNs and brain responses or behaviour?

Can the fit between DNNs and biological systems be improved by infusing DNNs directly with neural or behavioural data? Will such methods yield DNN models that more closely resemble the human brain and better predict successes and errors in human behaviour?

Given the diversity of models and modelling goals in cognitive science, how can we ensure fair assessment in terms of explanatory power? How can we unify standards for reporting and model use in an interdisciplinary research field where participating disciplines have disparate goals and standards?

Can we practically overcome the barriers for the clinical use of DNNs, for example, in patients with cortical damage (e.g., after stroke or after surgery)? In such cases, can DNNs be used as substitution devices that sufficiently restore cortical functioning (e.g., in vision loss)?

How can we ensure that cognitive and applied research using DNNs conforms to ethical norms such as privacy, identity, agency, and equality, and how do we deal with deviations from these norms? For example, how do we assess and deal with the challenge that might come about when a DNN trained on human data exhibits discriminative biases for gender or race?

In the future, DNNs may be widely adopted to replace human judgement, such as during job interviews, when a bank assesses a loan application, or in court. When used in such applied contexts, how do DNN's explanations and interpretability relate to issues of accountability, law, and legislation?

Acknowledgments

We thank Aude Oliva and Philippe Schyns for thoughtful comments and feedback. This work was supported by Deutsche Forschungsgemeinschaft (DFG) grants awarded to R.M.C. (CI241/1-1, CI241/3-1) and D.K. (KA4683/2-1).

References

1. Rumelhart, D.E. *et al.* (1987) *Parallel Distributed Processing*, A Bradford Book
2. McCulloch, W.S. and Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133
3. Krizhevsky, A. *et al.* (2012) Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.* 1, 1097–1105
4. Graves, A. *et al.* (2013) Speech recognition with deep recurrent neural networks. *IEEE International Conference on Acoustics, Speech and Signal Processing 2013*, 6645–6649
5. Mnih, V. *et al.* (2015) Human-level control through deep reinforcement learning. *Nature* 518, 529–533
6. Silver, D. *et al.* (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489

7. Graves, A. *et al.* (2016) Hybrid computing using a neural network with dynamic external memory. *Nature* 538, 471–476
8. LeCun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444
9. Hinton, G.E. (2007) Learning multiple layers of representation. *Trends Cogn. Sci.* 11, 428–434
10. Kubilius, J. *et al.* (2016) Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol.* 12, e1004896
11. Peterson, J.C. *et al.* (2017) Adapting deep network features to capture psychological representations: an abridged report. In *Proceedings of the Twenty-Sixth International Joint Conference of Artificial Intelligence*, pp. 4934–4938, Best Sister Conferences, Melbourne
12. Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* 10, e1003915
13. Cichy, R.M. *et al.* (2016) Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* 6, 27755
14. Horikawa, T. and Kamitani, Y. (2017) Generic decoding of seen and imagined objects using hierarchical visual features. *Nat. Commun.* 8, 15037
15. Yamins, D.L.K. *et al.* (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.* 111, 8619–8624
16. Eickenberg, M. *et al.* (2017) Seeing it all: convolutional network layers map the function of the human visual system. *Neuroimage* 152, 184–194
17. Güçlü, U. and van Gerven, M.A.J. (2015) Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35, 10005–10014
18. Cadena, S.A. *et al.* (2017) Deep convolutional models improve predictions of macaque V1 responses to natural images. *bioRxiv* Published online November 5, 2018. http://dx.doi.org/10.1101/201764
19. Gelfert, A. (2016) *How to Do Science with Models: A Philosophical Primer,* Springer International Publishing
20. Hempel, C.G. (2019) Explanation in science and in history. In *Frontiers of Science and Philosophy* (Colodny, R., ed.), pp. 7–33, Allen & Unwin Ltd.
21. Gelfert, A. (2016) Exploratory uses of scientific models. In *How to Do Science with Models: A Philosophical Primer* (Gelfert, A., ed.), pp. 71–99, Springer International Publishing
22. Gelfert, A. (2016) Strategies and trade-offs in model-building. In *How to Do Science with Models: A Philosophical Primer* (Gelfert, A., ed.), pp. 43–70, Springer International Publishing
23. Levins, R. (1966) The strategy of model building in population biology. *Am. Sci.* 54, 421–431
24. Morrison, M. and Morgan, M.S. (1999) Models as mediating instruments. In *Models as Mediators: Perspectives on Natural and Social Science* (Morgan, M.S. and Morrison, M., eds), pp. 10–37, Cambridge University Press
25. Matthewson, J. (2011) Trade-offs in model-building: a more target-oriented approach. *Stud. Hist. Philos. Sci. Part A* 42, 324–333
26. Ising, E. (1925) Beitrag zur theorie des ferromagnetismus. *Z. Phys.* 31, 253–258
27. Baddeley, A. (2011) Working memory: theories, models, and controversies. *Annu. Rev. Psychol.* 63, 1–29
28. Anderson, J.R. (1990) *The Adaptive Character of Thought,* Psychology Press
29. van Gelder, T. (1998) The dynamical hypothesis in cognitive science. *Behav. Brain Sci.* 21, 615–628
30. Kording, K. *et al.* (2018) Appreciating diversity of goals in computational neuroscience. *OSF Prepr.* http://dx.doi.org/10.31219/osf.io/3vy69 September 30
31. Gelfert, A. (2016) Between theory and phenomena: what are scientific models? In *How to Do Science with Models: A Philosophical Primer* (Gelfert, A., ed.), pp. 1–24, Springer International Publishing
32. Cartwright, N. (1999) Models and the limits of theory: quantum hamiltonians and the BCS model of superconductivity. In *Models as Mediators, Vol. 1* (Morgan, M.S. and Morrison, M., eds), pp. 241–281, Cambridge University Press
33. Boon, M. and Knuuttila, T. (2009) Models as epistemic tools in engineering sciences: a pragmatic approach. In *Philosophy of Technology and Engineering Sciences* (Meijers, A., ed.), pp. 687–719, Elsevier Science
34. Rajalingham, R. *et al.* (2018) Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* 38, 7255–7269
35. Hong, H. *et al.* (2016) Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* 19, 613–622
36. Yamins, D.L.K. and DiCarlo, J.J. (2016) Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365
37. Walker, E.Y. *et al.* (2018) Inception in visual cortex: *in vivo-silico* loops reveal most exciting images. *bioRxiv* Published online December 28, 2018. http://dx.doi.org/10.1101/506956
38. Bashivan, P. *et al.* (2018) Neural population control via deep image synthesis. *bioRxiv* Published online November 4, 2018. http://dx.doi.org/10.1101/461525
39. Schrimpf, M. *et al.* (2018) Brain-Score: which artificial neural network for object recognition is most brain-like? *bioRxiv* Published online September 5, 2018. http://dx.doi.org/10.1101/407007
40. Wimsatt, W.C. (2007) False models as means to truer theories. In *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality* (Wimsatt, W.C., ed.), pp. 94–132, Harvard University Press
41. Breiman, L. (2001) Statistical modeling: the two cultures. *Stat. Sci.* 16, 199–231
42. Yarkoni, T. and Westfall, J. (2017) Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12, 1100–1122
43. Achinstein, P. (1971) *Concepts of Science: A Philosophical Analysis,* The Johns Hopkins University Press
44. Black, M. (1962) *Models and Metaphors: Studies in Language and Philosophy,* Cornell University Press
45. Kay, K.N. (2017) Principles for models of neural information processing. *Neuroimage* 180, 101–109
46. Riesenhuber, M. and Poggio, T. (1999) Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025
47. Marr, D. (2010) *Vision,* MIT Press
48. Marblestone, A.H. *et al.* (2016) Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.* 10, 94
49. Kietzmann, T.C. *et al.* (2017) Deep neural networks in computational neuroscience. *bioRxiv* http://dx.doi.org/10.1101/133504
50. Knuuttila, T. (2011) Modelling and representing: an artefactual approach to model-based representation. *Stud. Hist. Philos. Sci. A* 42, 262–271
51. Samek, W. *et al.* (2017) Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arxiv.org/abs/1708.08296
52. Zhou, B. *et al.* (2015) Object detectors emerge in deep scene CNNs. *Int. Conf. Learn. Represent.* 2015
53. van der, Maaten L. and Hinton, G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605
54. Mahendran, A. and Vedaldi, A. (2014) Understanding deep image representations by inverting them. arxiv.org/abs/1412.0035
55. Zeiler, M.D. and Fergus, R. (2013) Visualizing and understanding convolutional networks. arxiv.org/abs/1311.2901
56. Yosinski, J. *et al.* (2015) Understanding neural networks through deep visualization. arxiv.org/abs/1506.06579

57. Simonyan, K. *et al*. (2013) Deep inside convolutional networks: visualising image classification models and saliency maps. arxiv.org/abs/1312.6034
58. Mordvintsev, A. *et al.* (2015) Inceptionism: going deeper into neural networks. *Google Res. Blog*
59. Zhou, B. *et al*. Interpreting deep visual representations via network dissection. arxiv.org/abs/1711.05611
60. Girshick, R. *et al.* (2016) Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 142–158
61. Xu, T. *et al*. (2018) Deeper interpretability of deep networks. arxiv.org/abs/1811.07807
62. Scholte, H.S. (2017) Fantastic DNimals and where to find them. *Neuroimage* 180, 112–113
63. Kriegeskorte, N. (2015) Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* 1, 417–446
64. Steinle, F. (1997) Entering new fields: exploratory uses of experimentation. *Philos. Sci.* 64, S65–S74
65. Burian, R. (1997) Exploratory experimentation and the role of histochemical techniques in the work of Jean Brachet, 1938-1952. *Hist. Philos. Life Sci.* 19, 27–45
66. Kisiel, T. (1973) Scientific discovery: logical, psychological, or hermeneutical? In *Explorations in Phenomenology: Papers of the Society for Phenomenology and Existential Philosophy* (Carr, D. and Casey, E.S., eds), pp. 263–284, Springer Netherlands
67. Hesse, M.B. (1963) *Models and Analogies in Science,* Sheed and Ward
68. Nayebi, A. *et al*. (2018) Task-driven convolutional recurrent models of the visual system. arxiv.org/abs/1807.00053
69. Rajaei, K. *et al.* (2018) Beyond core object recognition: recurrent processes account for object recognition under occlusion. *bioRxiv* Published online April 17, 2018. http://dx.doi.org/10.1101/302034
70. Tang, H. *et al.* (2018) Recurrent computations for visual pattern completion. *Proc. Natl. Acad. Sci. U. S. A.* 115, 8835–8840
71. Feest, U. (2012) Exploratory experiments, concept formation, and theory construction in psychology. In *Scientific Concepts and Investigative Practice* (Feest, U. and Steinle, F., eds), pp. 167–189, De Gruyter
72. Sterrett, S.G. (2014) The morals of model-making. *Stud. Hist. Philos. Sci. A* 46, 31–45
73. Waters, C.K. (2007) The nature and context of exploratory experimentation: an introduction to three case studies of exploratory research. *Hist. Philos. Life Sci.* 29, 275–284
74. Yosinski, J. *et al*. (2014) How transferable are features in deep neural networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Canada, Volume 2, 3320–3328
75. Razavian, A.S. *et al*. (2014) CNN Features off-the-shelf: an astounding baseline for recognition. arxiv.org/abs/1403.6382
76. Donahue, J. *et al*. (2013) DeCAF: a deep convolutional activation feature for generic visual recognition. arxiv.org/abs/1310.1531
77. Oquab, M. *et al*. (2014) Learning and transferring mid-level image representations using convolutional neural networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 1717–1724
78. Kümmerer, M. *et al*. (2014) Deep Gaze I: boosting saliency prediction with feature maps trained on ImageNet. arxiv.org/abs/1411.1045
79. Kriegeskorte, N. and Douglas, P.K. (2018) Cognitive computational neuroscience. *Nat. Neurosci.* 21, 1148–1160
80. Marcus, G. (2018) Deep learning: a critical appraisal. arxiv.org/abs/1801.00631
81. Hassabis, D. *et al.* (2017) Neuroscience-inspired artificial intelligence. *Neuron* 95, 245–258
82. van Gerven, M. (2017) Computational foundations of natural intelligence. *Front. Comput. Neurosci.* 11, 112
83. Lipton, Z.C. (2016) The mythos of model interpretability. arxiv.org/abs/1606.03490
84. Kubilius, J. (2017) Predict, then simplify. *Neuroimage* 180, 110–111
85. Woodward, J. (2017) Scientific explanation. In *The Stanford Encyclopedia of Philosophy* (Zalta, E.N., ed.), Stanford University
86. Skow, B. (2016) Scientific explanation. In *The Oxford Handbook of Philosophy of Science* (Humpreys, P., ed.), pp. 524–543, Oxford University Press
87. Hempel, C.G. and Oppenheim, P. (1948) Studies in the logic of explanation. *Philos. Sci.* 15, 135–175
88. Hempel, C.G. (1965) *Aspects of Scientific Explanation: And Other Essays in the Philosophy of Science,* Free Press
89. Salmon, W.C. (1984) *Scientific Explanation and the Causal Structure of the World,* Princeton University Press
90. Friedman, M. (1974) Explanation and scientific understanding. *J. Philos.* 71, 5–19
91. Kitcher, P. (1989) Explanatory unification and the causal structure of the world. In *Scientific Explanation* (Kitcher, P. and Salmon, W., eds), pp. 410–505, University of Minnesota Press
92. Fraassen, B.V. (1980) *The Scientific Image,* Oxford University Press
93. Naselaris, T. *et al.* (2018) Cognitive computational neuroscience: a new conference for an emerging discipline. *Trends Cogn. Sci.* 22, 365–367