nature reviews neuroscience

https://doi.org/10.1038/s41583-023-00705-w

Perspective

Check for updates

The neuroconnectionist research programme

Adrien Doerig[1,2,15] ✉, Rowan P. Sommers[3,15], Katja Seeliger[4], Blake Richards[5,6,7,8,9], Jenann Ismael[10], Grace W. Lindsay[11], Konrad P. Kording[9,12], Talia Konkle[13], Marcel A. J. van Gerven[2], Nikolaus Kriegeskorte[14] & Tim C. Kietzmann[1]

Abstract

Artificial neural networks (ANNs) inspired by biology are beginning to be widely used to model behavioural and neural data, an approach we call 'neuroconnectionism'. ANNs have been not only lauded as the current best models of information processing in the brain but also criticized for failing to account for basic cognitive functions. In this Perspective article, we propose that arguing about the successes and failures of a restricted set of current ANNs is the wrong approach to assess the promise of neuroconnectionism for brain science. Instead, we take inspiration from the philosophy of science, and in particular from Lakatos, who showed that the core of a scientific research programme is often not directly falsifiable but should be assessed by its capacity to generate novel insights. Following this view, we present neuroconnectionism as a general research programme centred around ANNs as a computational language for expressing falsifiable theories about brain computation. We describe the core of the programme, the underlying computational framework and its tools for testing specific neuroscientific hypotheses and deriving novel understanding. Taking a longitudinal view, we review past and present neuroconnectionist projects and their responses to challenges and argue that the research programme is highly progressive, generating new and otherwise unreachable insights into the workings of the brain.

Sections

[1]Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany. [2]Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands. [3]Department of Neurobiology of Language, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. [4]Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany. [5]Department of Neurology and Neurosurgery, McGill University, Montréal, QC, Canada. [6]School of Computer Science, McGill University, Montréal, QC, Canada. [7]Mila, Montréal, QC, Canada. [8]Montréal Neurological Institute, Montréal, QC, Canada. [9]Learning in Machines and Brains Program, CIFAR, Toronto, ON, Canada. [10]Johns Hopkins University, Baltimore, MD, USA. [11]New York University, New York, NY, USA. [12]Bioengineering, Neuroscience, University of Pennsylvania, Pennsylvania, PA, USA. [13]Harvard University, Cambridge, MA, USA. [14]Zuckerman Institute, Columbia University, New York, NY, USA. [15]These authors contributed equally: Adrien Doerig, Rowan P. Sommers. ✉e-mail: adoerig@uni-osnabrueck.de

Introduction

Although the study of cognition is a millennia-old endeavour (for example, already present in Aristotle's *De Anima*), the past decade has seen remarkable advances in both experimental and computational analysis techniques, yielding more powerful ways to study and model computations in the brain[1]. Yet, the level of abstraction at which cognition should best be understood remains a hotly debated topic. Modelling biology by replicating every molecular detail might not guarantee a deeper understanding of the core principles of cognition, any more than the brain of one person can serve as an explanation for the brain of another. Instead, the task of cognitive computational neuroscience is to find the right level, with enough fidelity to biology to preserve the essential mechanisms, but abstract enough to discard details not required for cognitive function, which reproduces the trajectory from actively sensed input, through internal representations realized in neural processes, to complex goal-directed behaviours[2].

Traditional experimental approaches often operate at the rather coarse-grained explanatory level of contrasting experimental conditions. For instance, by running visual neuroscience experiments with highly controlled stimuli, neural firing rates have been interpreted in terms of category selectivity: neurons are deemed selective for 'faces', 'houses' or 'tools'[3–6]. This approach has merit. Its controlled settings allow for maximal interpretability and suggest a clear taxonomy of neural selectivity[7]. Yet, human-interpretable labels for neural activity are limited by the imagination of researchers, or simply by language. But natural mechanisms are not necessarily bounded within these constraints: neural selectivity can often rely on more complex features that only imperfectly map onto human-interpretable categories[8]. In these cases, interpretable models are needed to find these non-interpretable complex features. In addition, showing selectivity for a high-level category, such as faces or houses, does not explain how the brain computes the representation from noisy sensory data, or what role category selectivity has in downstream function.

Together, these observations highlight the need for neuro-computational models grounded in sensory data that can bridge explanatory and computational levels and predict neural data and behaviour to answer the central questions of cognitive neuroscience: how can sensory input be linked to neural data across brain regions, not only at the level of individual cells but also at the population level? How can neural processes be linked to behaviour? How do neural representations change, not only through space but also through time (from fast synaptic adaptations and recurrent dynamics, through medium-term learning of a task, to much longer developmental trajectories)? How can past experience be encoded in the brain and which types of feature selectivity allow for task-general robust performance?

In short, for a complete picture of how cognition emerges, brain science needs interpretable computational models that go beyond the limits of human-interpretable labels for neural activity, that are applicable in naturalistic settings by being grounded in sensory data and that tie together multiple levels of explanation. Several characteristics of artificial neural networks[9] (ANNs) make them a model class well suited to tackle these challenges, with deep neural networks[10,11] particularly applicable. First, ANNs are made of simple units that collectively implement complex computations that drive the behaviour of the network. That is, they offer a framework spanning the single unit, collective dynamics, behavioural and computational levels. Second, ANNs use millions (sometimes billions) of synaptic parameters to encode rich domain knowledge while learning by optimizing connectivity over time. Third, ANNs are grounded in sensory input, which means that they can be trained on raw 'sensory' data to fulfil 'behavioural' needs, without the need for human-engineered input features, offering a link among sensation, cognition and action. Finally, by allowing for the comparison of different biologically inspired learning rules and objectives and how they interact with architectural network features, ANNs can help uncover how learning and cognitive development are made possible. Importantly, the architectural flexibility of ANNs and the different ways in which they can be trained allow for explorations of which biological details are needed for a given cognitive phenomenon. That is, by comparing ANNs that implement different biological details, researchers can test hypotheses about the computational effects of these biological features, much like traditional experimentation.

These characteristics allow researchers to rigorously test ANNs against large-scale behavioural and neural experimental data sets collected from a large array of brain regions[12–16] and to adjust the level of biological detail where needed — an approach markedly different from machine-learning engineering geared towards high performance on a small number of benchmarks[17]. Furthermore, the more recent focus on multilevel understanding of brain function of this approach goes beyond classic connectionist models of the twentieth century[18], which were limited to smaller networks to explain higher-level cognitive tasks without seeking explicit mappings from network units to the brain. Owing to the close integration into neuroscience, both in terms of network design and mapping of internal representations to brain function and neural data, we term this new approach 'neuroconnectionism' — a cohesive large-scale research programme centred around ANNs as a computational language for expressing falsifiable theories and hypotheses about multilevelled brain computation (Fig. 1).

Neuroconnectionism has already been successfully applied in a wide variety of neuroscientific settings, including vision[19–25], audition[26,27], semantics[28–31], language[32,33], reading[34], decision-making[35–38], attention[39], memory[40], game playing[41], motor control[42–45] and the formation and coding principles of brain areas[46–51] (reviewed elsewhere[52–58], further demonstrating that the larger neuroconnectionism research community involves, among many other areas, sensory

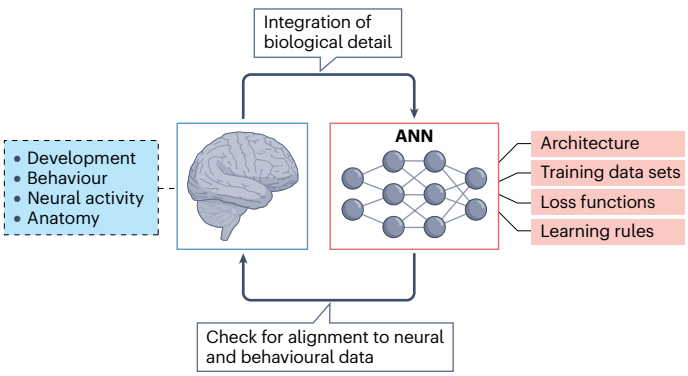


Fig. 1 | The neuroconnectionist research cycle. The integration of biological detail from neural and behavioural data across multiple scales informs the creation of new artificial neural network (ANN) models with different components, which are then tested for alignment with neural and/or behavioural data (left), leading to further cycles of model creation and model testing. Model creation involves four central ingredients: objective functions, training data sets, learning rules and architectures. Model evaluation involves hypothesis testing via tools such as representational similarity analysis, encoding models, comparisons of diagnostic readouts to human responses and in silico experimentation.

processing — in particular vision, our predominant research domain — language processing, memory, (meta) learning as well as movement and embodiment or robotics). Such developments across diverse research areas, the novel analytical tools that using large-scale ANNs provide, and the new possibilities that arise from this multidisciplinary modelling perspective are the reasons for substantial excitement in the scientific community.

At the same time, neuroconnectionism does not remain unchallenged. It has been raised that ANNs differ strongly from biology and that they often behave in non-human ways and that the complexity of the models prohibits true insights into brain function[59–62]. These challenges can be interpreted as suggesting that ANNs are not useful models for learning about the brain.

The aforementioned differences in perspective, as well as the increasing popularity of ANNs in neuroscience and beyond, demonstrate the crucial need for a clear conceptual understanding of the goals, tools, current empirical validity and future promises of neuroconnectionism. Although few researchers believe that ANNs should be abandoned entirely, the literature on the merits and shortcomings of ANN models[52,55,56,59–61,63–70], as well as discussions at conferences and on social media, commonly escalate into dichotomous debates. As a lack of clarity regarding the rationale and aims behind neuroconnectionism might be amplifying this binary discourse, a framing of the research programme in the philosophy of science will benefit the wider research community.

Hence, in this Perspective article, we provide such a presentation by introducing and evaluating neuroconnectionism as a Lakatosian research programme (Box 1). We comprehensively discuss the core rationale behind using ANNs in brain science and how to evaluate models against biological data. Next, we demonstrate how to derive new insights and understanding from ANNs and examine how to assess the promise of this approach. Through an explicit focus on clarifying the rationale and aims behind neuroconnectionism using the underlying philosophical framework, our Perspective article also extends beyond previous reviews of the merits and pitfalls of ANNs in brain science[52,55,56,59–61,64–68]. Our aim is to clarify the role of ANNs in neuroscience and provide the needed conceptual tools to help resolve some of the less-productive debates surrounding this emerging field.

Neuroconnectionism as a Lakatosian research programme

In the philosophy of science, Lakatos[71] proposed a general framework to evaluate scientific approaches. According to his view, science is typically carried out within research programmes. Such programmes share a hard 'core' of background assumptions that are not typically challenged from within the programme and contain a 'belt' of auxiliary hypotheses that are experimentally tested. Although the core cannot be altered without abandoning the research programme, the auxiliary hypotheses comprising the belt are (and should be) subject to change.

Given these two elements, core and belt, the value of a research programme is determined not just by its current experimental success relative to other research programmes but also based on whether it is progressive rather than degenerating — an explicitly longitudinal perspective (Fig. 2). The cores of progressive research programmes generate new theoretical insights and novel predictions in their belts, some of which are corroborated by empirical findings. This new knowledge advances the research programme by leading to further insights and testable hypotheses. Degenerating research programmes do not have these two characteristics: they often lack new theoretical developments and novel predictions to be tested, instead devolving into repeated corroboration of very similar ideas. In summary, in a progressive 'successful' research programme, background assumptions in the core help researchers to generate new knowledge and testable hypotheses in the belt, whereas in a degenerative 'failing' research programme, background assumptions in the core lead to stagnation.

The present discussion does not depend on accepting the details of the philosophy of Lakatos. What matters most is that scientific theories always have core guiding principles that are relatively isolated from direct empirical testing — an uncontroversial view in the philosophy of science (Box 1). These core-guiding principles define directions of inquiry, including which experiments are conducted, and the type of empirical results needed to corroborate or weaken theories. Hence, all scientific theories are judged through a holistic assessment of their successes and their fertility in guiding experimental pursuits. The Lakatosian perspective that we have adopted is merely one helpful way of expressing these general ideas so we can apply them to neuroconnectionism as a research programme mentioned subsequently (Fig. 2).

The neuroconnectionist core

To lay out the core of neuroconnectionism, the modelling aims laid out in the introduction can be summarized as a (non-exhaustive) list of desiderata. Thus, a good model of brain computations underlying cognition should:

a. Specify which computations are carried out by the brain (computational level).
b. Show how these computations lead to complex behavioural patterns that can be tested in experiments (behavioural level).
c. Show how these computations lead to complex neural dynamics that can be tested in experiments (single unit level and collective dynamics level).
d. Show how these computations can be carried out in complex naturalistic settings, beyond simplified highly controlled experiments (rich domain knowledge).
e. Show how these computations can be grounded in sensory information, rather than high-level features provided by human-interpretable labels (sensory grounding).
f. Show how these computations arise from adaptive processes that unfold at multiple timescales (from processing dynamics to developmental trajectories).

Simple models with a small number of directly interpretable parameters are not ideal candidates because they cannot achieve each desideratum. They are incapable of dealing with naturalistic settings because they lack sensory grounding (desideratum d) and rich domain knowledge (desideratum e), both of which require complex computations that can only be achieved in highly parameterized models. In addition, a multilevel (desiderata a–c) and dynamic (desideratum f) understanding of cognition that spans from neurons to behaviour most likely requires models with distributed and iterative computations, because that is how real neural networks operate. Hence, the complex, distributed and iterative computations underlying cognition in the brain probably can only be modelled using other complex, distributed and iterative processes, which are necessarily highly parameterized.

However, complex models are not completely ideal candidates either. First, models need to be computationally tractable (runnable at scale on current computers). Second, millions of parameters need to be adequately tuned to encode domain knowledge for complex and sensory grounded behaviour. As this is impossible to do by hand,

training algorithms are required to do this automatically. Although ANNs trained via deep learning optimizers can achieve this, the high number of parameters needed renders the interpretation of individual units more challenging, which then requires additional methods to establish the desired mapping among model, brain data and behaviour (see the next section 'From core to belt: the neuroconnectionist toolbox' for complete details). Taken together, which model class is powerful enough to accomplish the difficult task of brain modelling, while still being computationally tractable and interpretable enough to yield true insights for brain science, becomes a central question.

To the proponents of neuroconnectionism, ANNs achieve this intricate balance. They are sufficiently abstract to be tractable and

Box 1

Theory selection and philosophy of science

Through much of the twentieth century, the Popperian[299] view that theories are rejected when they are falsified in tests was dominant. A scientific theory generates predictions and tests are run to see whether the predictions are correct. If they are not, the theory is falsified and can be rejected (it may take more than one test to show that the failure was not experimental error, but once the result is accepted, the theory must be rejected).

The Popperian view assumes that the logic of disconfirmation follows the following schema:

If T, then O

Not O

Hence, not T

Where T is a theory and O is an observation.

It was noticed first by Pierre Duhem (the French theoretical physicist and historian of science) in 1906[300,301], and later reinforced by Quine[285] (one of the most influential analytic philosophers in the twentieth century) that science does not work like that in practice and could not work like that in principle. One typically (and perhaps always) needs to combine the hypothesis to test with auxiliary beliefs to extract empirical predictions. When it is made fully explicit, the logic looks as follows:

If T, and A1, and A2, and A3,..., and An, then O

Not O

Hence, not T, or not A1, or not A2, or not A3,..., or not A*n*

Where (A1, and A2, and A3,..., and A*n*) are auxiliary hypotheses needed to generate predictions.

For this reason, it is never a single hypothesis, but a whole collection of hypotheses that generates predictions, any one of which might be at fault if the prediction is not vindicated.

This fact about the logic of confirmation has been one of the centrepieces of twentieth-century philosophy of science[302]. Quine[303] popularized the idea that a theory forms a web of beliefs related by inferential connections. The web has a topology with beliefs that can be most directly subjected to empirical tests at the periphery, and others that are insulated from direct testing by long chains of intermediary hypotheses. Beliefs at the periphery describe localized observable facts; beliefs at the centre describe the kinds of general beliefs that guide the explanation of a whole body of phenomena. These general beliefs are highly connected in the web and separated from empirical predictions by mediating propositions. Thus, one could hold onto the general beliefs in the face of mounting evidence if one was willing to adjust other more peripheral parts of the web.

The distinction of Lakatos between core and belt acknowledges this holistic nature of confirmation, while offering testing pragmatics well suited to capture how science actually works. In the Popperian view, testing is a matter of checking whether a theory accords with fact. In the Lakatosian view, testing goes hand in hand with the development of theory as one adopts a set of core principles as a kind of working hypothesis, proceeding on the assumption that they are correct and using them to try to understand the phenomena. Testing is a process that involves striving to bring theory and fact into closer agreement by exploring ways in which the core principles can be preserved while accommodating the accumulating evidence. If the core principles are held fixed and leeway is granted to explore alternative auxiliary hypotheses, testing can be directed at the belt, giving the theory every chance of preserving the core while accommodating disconfirmatory evidence. In this process, the core is not rejected when disconfirmatory evidence is found — instead, new studies are conducted to find out whether the disconfirmatory evidence can be explained as a failure of auxiliary hypotheses (for example, some measurement error). A theory is rejected not as the result of a direct conflict with the evidence, but because the attempt to preserve the core principles becomes so cumbersome that they cease to form a productive working hypothesis for continued testing and the discovery of new insights.

For example, in the field of astronomy, deviations in planetary trajectories from the smooth ellipses predicted by Newtonian mechanics were observed. Instead of rejecting Newtonian laws owing to these challenging empirical data, scientists assumed the correctness of the laws and tested auxiliary hypotheses (such as the presence of an unseen planet) that might explain the orbital deviations. Hence, a belt claim was falsified (the number of planets in the solar system) but the core was not abandoned (Newtonian mechanics). The core is changed only when it becomes unproductive to hold onto it, because it no longer leads to new hypotheses or because its hypotheses are not corroborated. Changing cores in effect involves changing research programmes and is therefore similar to a Kuhnian scientific paradigm shift[304], leading to a complete overhaul of theories and the language they use to describe the world. For example, in the twentieth century, evidence accumulated against Newtonian celestial mechanics that could not be solved assuming the correctness of the laws, which led to its rejection and the development of general relativity, a novel progressive core that changed the way the universe is thought about and led to great discoveries, such as curved space–time and black holes, and technological applications, such as more precise space-travel, astronomical and GPS tools.

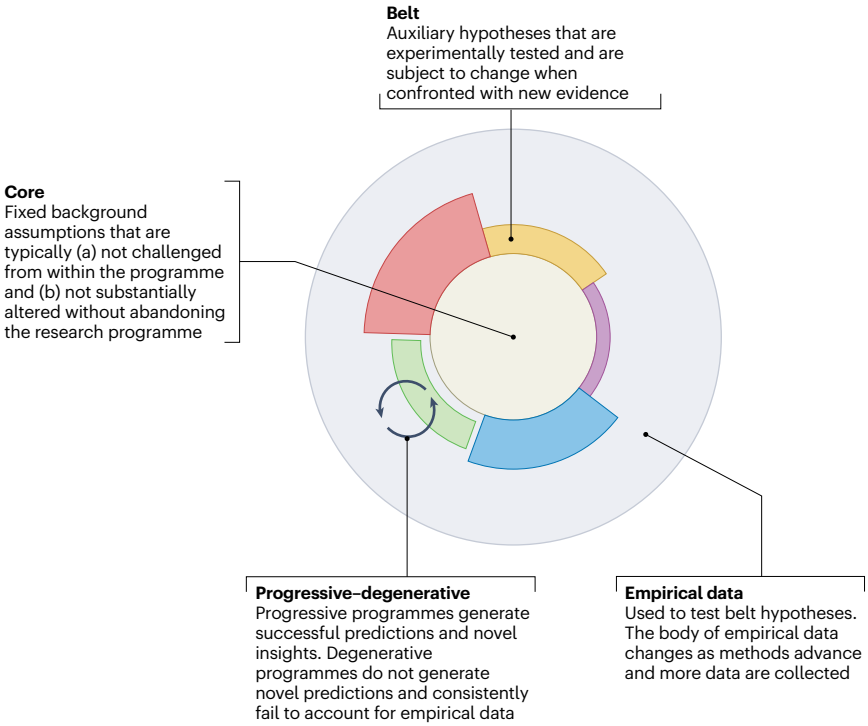


Fig. 2 | Lakatosian research programmes. A conceptualization whereby research programmes are composed of a core of fixed background assumptions and a variable belt of auxiliary hypotheses. Empirical data are used to test and falsify belt hypotheses without changing the core. In the Lakatosian view, the entire research programme is not immediately falsified by conflicting empirical data. Instead, it is judged on its ability to successfully adapt its belt hypotheses to satisfy empirical constraints, which is indicated longitudinally by whether the research programme generates new insights and corroborates belt hypotheses (progressive) or not (degenerative).

trainable, but also retain sufficient biological detail in their algorithmic structure to map them onto neural and behavioural data. In other words, ANNs live in the Goldilocks zone of biological abstraction (Fig. 3), striking the required balance between biological realism and algorithmic clarity. By contrast, models with too much biological detail, such as in silico copies of brain regions, are not computationally tractable at the large scale required for sensory-grounded, behaviourally complex cognitive tasks. Therefore, such models cannot provide a connection from low-level neurons to higher-level cognitive function and thus fall outside this Goldilocks zone. Furthermore, unnecessary detail complicates understanding — abstraction is central to revealing which details matter. At the other end of the spectrum, models that are too distant from biology, such as classic box-and-arrow cognitive models[72], fall outside the Goldilocks zone because they are too abstract and cannot be linked to biology, nor grounded in sensory input. But ANNs strike the right balance by providing a level of abstraction much closer to biology but abstract enough to model behaviour: they can be trained to perform high-level cognitive tasks, while they simultaneously exhibit biological links in terms of their computational structure and in terms of predicting neural data across various levels — from firing rates of single cells, to population codes and on to behaviour.

In addition to being in this Goldilocks zone of computational abstraction, which makes ANNs tractable and mappable to biology, they allow for a productive research cycle of generating, implementing and testing hypotheses about brain computations. Indeed, ANNs are defined by their architecture, data set statistics, objectives and learning rules, which can be mapped onto central questions of brain science (see 'From core to belt: the neuroconnectionist toolbox'). This includes disentangling the interacting contributions of pre-specified structure (for ANNs: determined by the architecture) and experienced input (for ANNs: training data set), why neural selectivity in any given brain region is the way it is (for ANNs: which objectives are being optimized) and how the brain may adjust its internal representations (for ANNs: credit assignment or learning rules). All these questions can be studied across levels of explanation and temporal scales, incorporating rich domain knowledge grounded in sensory data, in line with desiderata (a–f). The resulting models can be tested with great precision, and advanced methods exist to derive understanding from models, making neuroconnectionism a useful computational language for thinking about and describing brain computations underlying cognition.

Thus, the Lakatosian core of the neuroconnectionism research programme consists of the following fixed background assumptions:

1. Brain science requires complex, distributed and iterative models to address desiderata (a–f).
2. ANNs offer a highly suitable computational language: sufficiently abstract to be computationally tractable and reproduce cognitive functions, while still being close enough to biology to relate to, implement and test neuroscientific hypotheses.

This motivates the neuroconnectionist claim that brain computations, representations, learning mechanisms and inductive biases are best understood via modelling in ANNs, defined by their architectures, training data sets, objective functions and learning rules, which must be tested against neural and behavioural data (Fig. 1).

Distinctive features of neuroconnectionist models

The desiderata and fixed background assumptions mentioned earlier aim to capture the core tenets shared by the many projects that contribute to the neuroconnectionism research programme. Although a rigid definition of a large-scale and highly dynamic research programme is bound to fail to capture its diversity, just like how members of a family do not share all traits, the respective individual projects possess a family

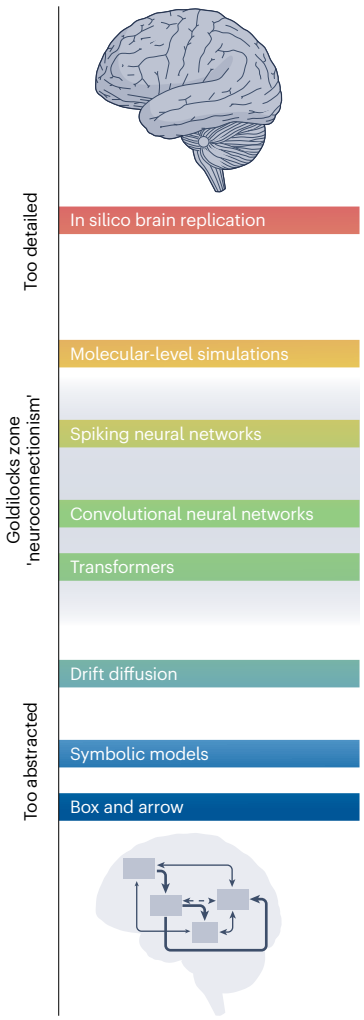


Fig. 3 | Schematic of the Goldilocks zone of biological abstraction. An analogy to convey the balance between detail and abstraction. This analogy is borrowed from astronomy, in which the search for inhabitable exoplanets means looking for planets orbiting at the 'right' distance from their stars to have liquid water. If they are too close, temperatures are too high and water evaporates. If they are too far, temperatures are too low and water freezes. The temperature has to be just right, as in the Goldilocks fairytale. Analogously, models that are too close to the biological brain fall outside the Goldilocks zone because they have too much biological detail: these models cannot be run or trained at scale to perform complex cognitive tasks from sensory grounded evidence. Models that are too abstract also fall outside the Goldilocks zone: these models can neither be easily linked to biology, nor be grounded in sensory input. As unnecessary detail complicates understanding, models need to focus on incorporating the biological elements crucial for explaining brain computation at an appropriate level of abstraction. Neuroconnectionism offers a coherent and computationally tractable framework for brain science in which models can vary in how much they abstract away from biology. This enables researchers to determine which biological details are needed in the models they create to test their hypotheses regarding the brain computations underpinning cognition.

resemblance[73]. Hence, the member projects of neuroconnectionism are best understood as constituting a diverse but cohesive family of approaches that share distinctive features.

Explicit mapping between ANNs and biology. Neuroconnectionist researchers seek explicit mappings between ANNs and the brain through hypothesis-driven research, which differs from engineering goals. Around the time of the first large-scale vision networks nearly a decade ago[74,75], neuroconnectionist models were borrowed directly from engineering applications. For example, engineering models that performed the best on 2014 engineering benchmarks were also better at predicting brain activity[25]. This correspondence between engineering and neuroscientific goals may have led to a form of 'computational opportunism', in which researchers could directly test machine-learning models against neuroscientific data without having well-formed hypotheses. Generally, fitting an engineering model to brain data without testing a biological hypothesis is not part of neuroconnectionism. Moreover, computational opportunism is no longer a valid strategy because, while more recent engineering architectures have better task performance, they currently have worse alignment with neural data[12]. Indeed, when taking the hierarchical nature of the visual system into account, the past 5 years have seen more and more limitations for ANNs borrowed directly from engineering in pursuing neuroscientific goals[76]. Together, focusing on engineering goals based on task performance alone is not sufficient for obtaining a neuroscientific understanding, highlighting the need for neuroconnectionism to develop its own models and metrics. Models strongly driven by engineering goals, such as generative adversarial networks or transformers, can only contribute to neuroconnectionism if they can be explicitly mapped to biology and are used to test hypotheses about brain computations[77–79].

Understanding via abstraction. Neuroconnectionist models are primarily aimed at explaining brain computations at a level of abstraction that links neurons directly to their functional relevance for the behaviour of the system, not aimed at describing biology with the highest possible detail. Biological detail is added to models as part of hypothesis testing, to see which details are necessary for explaining behavioural and neural data. This makes neuroconnectionism different from approaches aiming to perfectly replicate a human brain in silico[80] and from approaches using biophysical models[81] aiming to model every aspect of a neuron or neural circuit, as neither of these approaches typically adapts its level of abstraction to best model cognitive processes.

Distributed representations and computations. In ANNs, the modelled property emerges from the collective behaviour and dynamics of simple units, which, taken independently, do not exhibit the modelled property. This distributed nature of ANNs is central to neuroconnectionism as it readily bridges between explanatory levels, from single units through collective dynamics and onto behaviour, and requires special interpretation frameworks to cope with the distributed nature of ANN computations (see the neuroconnectionist toolbox section). By contrast, traditional models such as classic box-and-arrow models in cognitive neuroscience, as well as models equating a given brain region with a given cognitive function, or simpler computational models in which each parameter has an interpretable functional role, such as drift-diffusion models, or models based on signal detection theory do not rely on distributed computations. Symbolic rule-based approaches, such as Good Old-Fashioned Artificial Intelligence, where each variable, and each rule applied to variables, is designed to have a human-interpretable meaning, are also similarly distinguished from neuroconnectionist models.

Iterative training and inference. The behaviour and internal states of many distributed and iterative processes — even simple ones such as Conway's Game of Life — often cannot be predicted by simple non-iterative models but only by distributed and iterative models, as shown by mathematical proofs[82,83]. The only way to predict the evolution of these systems is to run the distributed and iterative process or run a similar distributed and iterative model system. Similarly, the complex and highly nonlinear computations of the brain likely cannot be simplified into easily interpretable models or equations either. Complex distributed and iterative models, such as ANNs, are likely required. However, the resulting high dimensionality of such models makes it impossible to tune all parameters by hand, which is why ANNs require iterative training by successively applying millions of weight updates to optimize one or many objectives (see the next section). Next, at the inference stage, when the trained ANN is used to infer the behaviour and dynamics, they cannot be simplified into a simple interpretable equation: to know what result the model predicts, one needs to run the model. The iterative nature of the training and inference stages distinguishes neuroconnectionism from approaches that can be formulated in a small number of non-iterative equations, such as (hierarchical) Bayesian models.

Edge cases. Although the features discussed earlier characterize neuroconnectionist models and distinguish them from approaches in contemporaneous research programmes to neuroconnectionism, edge cases exist. For example, although grounding in sensory input is a desideratum of the neuroconnectionist approach, not all current neuroconnectionist models are grounded in sensory input. For instance, in language or memory models, the input to the model often consists of high-level concepts, such as words. Still, we consider these models neuroconnectionist as the sensory nature of the inputs may not be relevant in these particular cases and could be added to later models if needed without a change in framework, for example by including a visual ANN as a front-end to deal with naturalistic visual stimuli. As another edge case example, some neuroconnectionist models might not explicitly use a 'behavioural' objective function, but rather attempt to operate under externally defined constraints, such as energy efficiency[84]. Moreover, models directly fitted to neural data, instead of being trained on a behavioural task, are considered neuroconnectionist because they provide ways of hypothesis testing needed to determine which architectures are generally capable of reproducing neural dynamics[23]. In addition, most models do not aim for explaining cognition in general, but rather focus on explaining specific components. Thus, understanding which aspects of cognition can only be modelled jointly is part of the hypothesis-testing process. For example, sensory–motor interactions in embodied models may or may not be required to explain certain aspects of visual processing.

Summary of the neuroconnectionist core. Motivated by desiderata (a–f), the core of the neuroconnectionist research programme is to use ANNs as a language for expressing computational neuroscientific hypotheses. This is possible because ANNs reside in a Goldilocks zone of computational abstraction, which allows them to model complex cognitive functions grounded in sensory data while still being mappable to biological features. Neuroconnectionist models form a loose but cohesive family, centred around the goal of implementing different biological details and testing which are needed to explain cognition. Although there are edge cases, an explicit mapping to biology, understanding via abstraction, distributed representations or computations as well as iterative training and inference are widely shared characteristics of neuroconnectionist models.

From core to belt: the neuroconnectionist toolbox

In the Lakatosian view, the core of a research programme is not directly falsifiable, but it is used to derive falsifiable belt hypotheses, which are tested against empirical data (Fig. 2). In the neuroconnectionist research programme, falsifiable belt hypotheses are evaluated by building, training and testing ANNs against neural and behavioural data (Fig. 1). To do so, we instantiate a neuroscientific hypothesis into the neuroconnectionist language using the neuroconnectionist toolbox described in this section. According to standard scientific practice, the resulting ANNs are contrasted with appropriate control models and evaluated on their ability to explain the neural and behavioural data related to the hypothesis in question. All of this requires an extensive toolbox for instantiating, testing and interpreting new models — steps which are discussed in turn next.

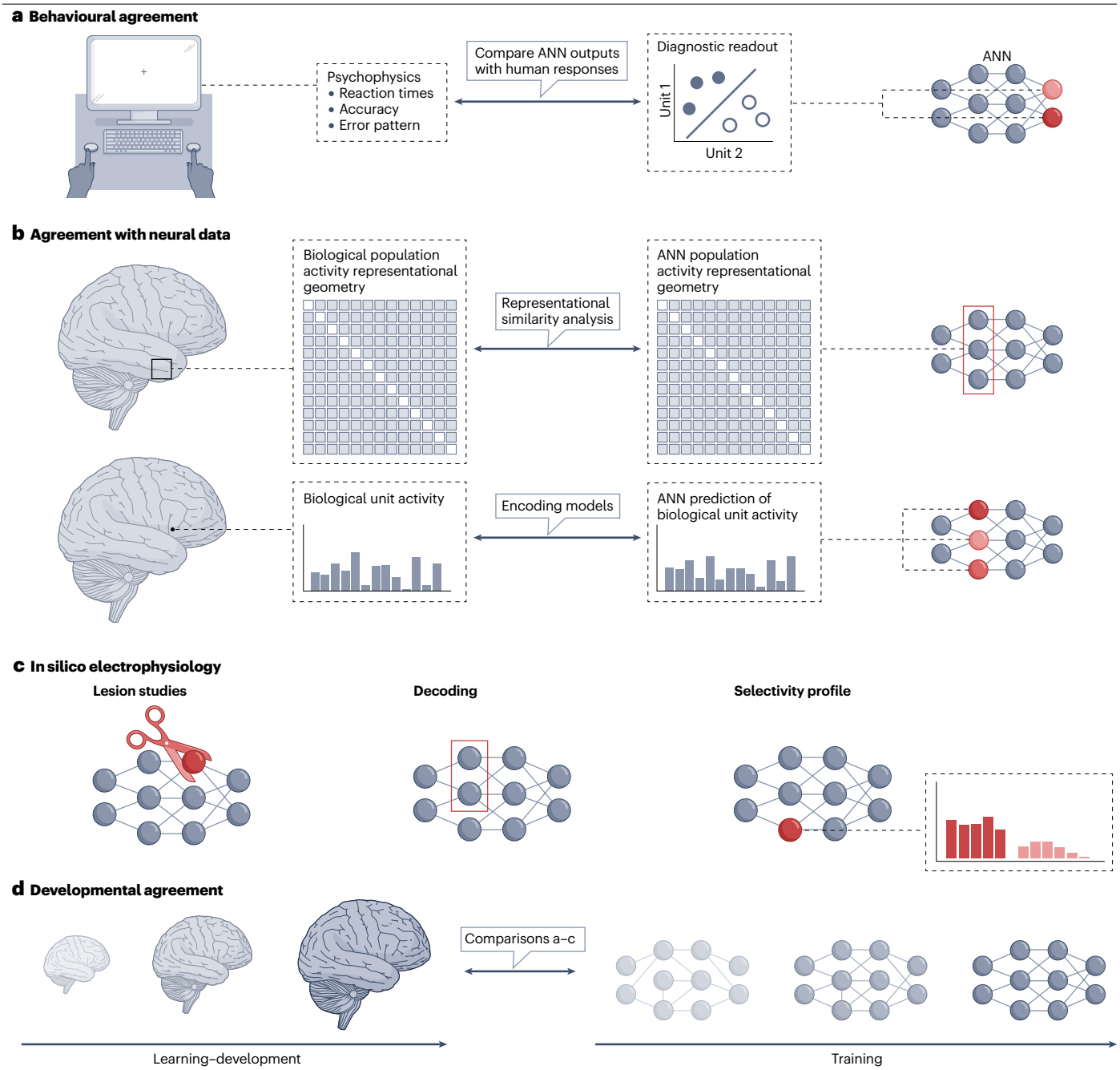
Model instantiation

ANNs are built and trained on the basis of four central ingredients: architectures, data sets, objectives and learning rules[52]. Each of these ingredients determines how and what the model learns, which in turn influences how well it matches biological data.

Architectures define a computational scaffold within which a given network can be expressed. Network architectural features include layers and computational unit types, which are abstracted from biology to various extents. Layer types include, among many others, random reservoirs[85,86], convolutional layers[20,87] and other more advanced designs. Common types of computational units range from very simple rectified linear units for summation, to more complex units modelling basic memory (for example, in long short-term memory networks[88]). Different architectures come with different inductive biases, which influence which functions can be learnt and thereby impact how well the resulting ANN compares with biological data. Therefore, research iterations on architectures are a central element of neuroconnectionism to determine which level of detail is needed to match biological data. For example, researchers have tested how integrating various architectural aspects inspired from biology, such as recurrent connectivity[23,89–95], richer rate-based neurons[96–98], spiking neurons[99–103] and neurons with multiple compartments[104,105], into network design impacts performance against neural and behavioural data.

Although architectural design has a role analogous to large-scale brain structure by determining the potentiality of the ANN, the individual network parameters still need to be optimized to perform a given task. Networks need to be trained on data sets to learn these parameters. Many large-scale data sets containing natural images, auditory signals or text corpora are openly available. Typically, external data sets — such as a collection of 'perceptual' inputs, sometimes associated with labels that the network must predict — are used for training. In addition, networks can be trained to learn parameters directly from brain activity[23,90,106–113], leveraging recent efforts to record large neural data sets[14,114–118]. As the training data set determines the input statistics from which network parameters are learnt, different data sets can lead to very different networks. Therefore, an important avenue of the research programme is to develop more naturalistic data sets (reviewed elsewhere[119]) and iterate over models trained on different data sets, testing what features of the data are required to match biological data.

One or several objectives — mathematically described by loss functions — determine what networks learn on the basis of the input



statistics determined by the data set. There are a large variety of objectives used, including supervised (classification and scene captioning), unsupervised (contrastive learning[120,121], predictive coding[122,123], image generation[50,124,125], temporal stability[126–130] and energy efficiency[84]) and behavioural reward[35,41,131]. Different objectives impact what the network learns and thereby the ability of the network to model different brain areas[132]. Therefore, an important aspect of neuroconnectionism is to iterate over models trained with different loss functions to test different hypotheses about the objectives of the brain. Ultimately, several low-level objectives may be subsumed in higher-level objectives, such as 'fitness'.

To minimize loss functions, the weights of the ANN must be updated, thus storing information about how to produce the desired output on the basis of sensory-grounded inputs into the connections of the network. To this end, a learning rule must change the weights of each individual unit in accordance with the contribution of the unit to the error of the whole network. How to attribute the contribution of each individual unit in the network to the overall network error is called the credit assignment problem[11]. Given that ANNs can have millions of network units — operating over extended time in the case of recurrent networks — creating learning rules is far from trivial. By far, the most common learning rule used is backpropagation[133]. However, backpropagation in its standard form is not biologically plausible, but multiple avenues of research seek to define backpropagation in biologically more plausible ways[105,134–138]. Other learning rules

Fig. 4 | The current neuroconnectionist toolkit for model testing. The neuroconnectionist toolkit contains many techniques that enable researchers to thoroughly evaluate models against neural and behavioural data. **a**, Behavioural agreement. Outputs of artificial neural networks (ANNs) are compared with human responses in diverse settings, such as classification of errors and accuracy, reaction times, action patterns, and others. **b**, Agreement with neural data. Presenting identical stimuli (input) to the brain and computational model, the recorded brain activity patterns are directly compared with ANN activity patterns. The most common methods are representational similarity analysis (comparing representational geometries of population activities) and encoding models (predicting brain activity from ANN units via linear regression). **c**, In silico electrophysiology. ANNs are studied as in silico models of cognitive functions with standard neuroscientific methods, such as manipulations and lesions, information decoding, unit-based tuning functions and others. Effects of design choices such as recurrent connections or manipulations such as detailed lesioning patterns can be studied extensively in this setup. These manipulations go beyond what is possible in vivo. **d**, Developmental agreement. Comparing different stages of training in ANNs with different stages of learning in biological brains permits insights into cognitive development. Examples include behavioural patterns, map formation or changes in neural selectivity with visual experience. All of these approaches can also be applied to compare ANNs with neural and behavioural data from non-human species (such as primates, rats and mice). Multiple alterations to ANNs are possible for each component part: objective functions (classification, energy efficiency, contrastive learning, agreement with neural data, action and slowness); training data sets (visual, auditory, text, motor signals, somatosensory and neural data); learning rules (gradient descent, Hebbian learning and evolutionary algorithms) and architecture (feedforward, recurrent, rate code or spiking units, locally or fully connected and convolutional). Multiple sources of data exist at each neural and behavioural scale: anatomy (diffusion tensor images, T1-weighted structural imaging, connectome, cortical layer architecture, cell morphology and cell types); neural activity (including functional magnetic resonance imaging, magnetoencephalography, electroencephalography, electrocorticography, array recordings of local field potentials and single-cell recordings); behaviour (accuracy, reaction times and error patterns from classic or naturalistic paradigms) and development (learning trajectories, curriculum learning evolutionary priors).

such as Hebbian learning[139], predictive coding[140] or self-organizing maps[141] exist too and each algorithm of learning rules has its own characteristics, benefits and limitations. On the one hand, gradient descent tends to learn the input features with most variance first[142] and produces efficient neural codes[143]. On the other hand, Hebbian learning is a simpler local rule that has been directly observed during learning in biological systems[144]. Typically, a single learning rule is used, but one could also combine several, for example, using backpropagation in combination with Hebbian learning. It is worth noting that each learning rule comes with its own hyperparameters, such as a learning rate and how the learning rate changes over training. As with the other ingredients used to cast coarse-grained neuroscientific hypotheses into mathematically precise neuroconnectionist belt hypotheses, iterating over learning rules is important to understand how different learning rules impact the match of networks to biological data.

In summary, the architecture, training data set, objective and learning rule can be easily manipulated by the experimenter to iterate complex task-performing ANNs, which can then be compared with biological data to test different neuroconnectionist belt hypotheses. Because these degrees of freedom are human-interpretable and orders of magnitude smaller than the number of parameters in the ANN, this provides a powerful, yet flexible, language for hypothesis testing and for generating new insights and predictions for the research programme (Fig. 1).

Model testing

Hypothesis testing in neuroconnectionism relies on testing trained ANNs against empirical data on various levels, from neural data up to behavioural patterns. Because of the flexibility of learning in ANNs, models that are structurally very different from the brain can nevertheless enable successful predictions of behavioural patterns, or of the neural activity in certain brain areas (a form of multiple realizability; for example, reviewed elsewhere[145,146]). Hence, to ensure that an ANN implements a given cognitive function in a similar way to the brain, its activities and output need to map onto brain processing across levels, from neural data to behavioural, ideally while considering physical constraints the brain faces such as metabolic costs[84,147–149] and wiring length[51]. Importantly, no single-model testing method is perfect and various complementary approaches are needed. Thus, developing good metrics to compare empirical data and ANNs across levels is a crucial part of neuroconnectionism.

Behavioural agreement. At the behavioural level, the outputs of ANNs can be compared in several settings, from detailed psychophysics to large-scale benchmarks (Fig. 4a). Coarse measures such as overall task performance on large benchmarks are useful but often fail to arbitrate between models, as multiple and different kinds of ANNs can reach human-level performance at tasks for which humans were until recently deemed the gold standard such as object recognition[74,150,151], board games[152] or video games[153,154]. To complement these coarse benchmark measures and help arbitrate between models, several more fine-grained methods exist. These include the use of diagnostic readouts to characterize the information represented in a population of units from the ANN and to then translate this information into behaviourally relevant measures such as reaction times[155], detailed analysis of error patterns[153], testing on out-of-distribution examples[156–158] and reproducing psychophysical results that target particular aspects of processing[89,159–165] (reviewed elsewhere[166–169] for specific discussion about how to compare human behaviour and ANN behavioural predictions). A unified model addressing years of psychophysical experimentation is an important target yet to be achieved by the research programme.

Neural data agreement. At the neural level, the activity patterns of ANNs can be compared with the brain in several ways including using representational similarity analysis (RSA)[170,171]. RSA characterizes the internal representations of a system by quantifying the dissimilarities between the population activity patterns during different experimental conditions (for example, the activity patterns to various stimuli), summarized in representational dissimilarity matrices (RDMs). Internal representations of ANNs and the brain are deemed similar if the corresponding response geometries agree (Fig. 4b). Thereby, RSA side-steps the problem of finding an explicit mapping from ANN units to individual neurons or voxels and focuses instead on population-level representational geometries. For ANNs, RDMs can be computed using population activity patterns from all units of a whole network, all units of a network layer, units in a feature map (which are selective for the same feature) or individual units. These ANN RDMs can then be directly compared with brain RDMs from neural populations or brain regions of interest, or an additional data fitting step can be integrated to optimize

the agreement between the ANN and the brain. A direct comparison of representational geometries between brain data and/or ANNs[119] has the benefit of not needing any free parameters to realize the ANN–brain mapping (thus avoiding the problem that parameter fitting might do most of the heavy lifting in matching brain data). A more flexible RSA approach combines multiple RDMs obtained from ANNs using linear reweighting to optimize the agreement with brain RDMs[22,172,173]. Although this allows for a better agreement between model and brain data, an important challenge for these reweighting approaches, or any analysis that gives parametric flexibility to the ANN–brain mapping, is that the flexibility can render invisible otherwise prominent differences among network candidates. Recent work has started addressing this issue by improving current RSA methods[174] and clarifying which aspects of brain computation should be targeted using RSA[175].

In addition to RSA, which is predominantly aimed at characterizing responses at the population activity level, encoding models can be used to predict the activity of single neurons or voxels across a range of conditions[94,176,177]. Here, the activity of each biological unit (neuron or voxel) is predicted as a linear combination of ANN unit activations. Hence, this is a mass univariate approach, in which each biological unit is predicted independently. To prevent overfitting, the underlying generalized linear models are typically regularized, with new methods for doing so in constant development[110,112]. One challenge is that units in encoding models are not constrained by which biological counterpart they can explain. For example, in principle, this lack of constraint implies that higher-level brain regions can be explained by lower-level network features, or that the activity of thousands of brain cells can be explained by the response of a single network unit with broadly similar selectivity. Moving towards more structured tests of the alignment between brain representations and ANNs using encoding models, new developments are underway to include an ordered hierarchical mapping from ANNs to brain regions[76,178].

The RSA and encoding model approaches are correlational. Building on encoding models, new techniques exist, which use ANNs to control a single target neuron or brain area[179–181]. For example, stimuli can be optimized to maximally drive neural activities in V4 by relying on ANNs to predict which activities will be evoked by different stimuli[179]. This provides a more casual approach to test the link between ANNs and brain processing.

In silico electrophysiology. In addition to estimating the level of agreement between ANNs and biological brains in terms of behaviour and neural recordings, ANNs themselves can be experimented on to better understand their inner workings. As all units, their activities and their connectivity are immediately accessible, almost any 'in silico' electrophysiology experiment is possible (Fig. 4c). In silico experiments are orders of magnitude faster to conduct than experiments on biological brains and are, for now, free of ethical concerns that come with classic experimentation. These in silico experiments include reliance on network initialization[182] and tests for the emergence of brain-like computations in individual network units[84], selectivity profiles[47,51,183] and cell types[184]. To achieve these ends, searchlight decoding, measures from signal detection theory, tuning curve analysis and many more standard neuroscientific methods can be applied. In addition, different parts of ANNs can be selectively lesioned to test their impact on the ability of the network to map onto brain function. For example, the effect of recurrent connections can be directly assessed by ablating them[23,89,162]. In silico lesion electrophysiology studies are not only limited to analyses of networks on their own but also can be used to evaluate changes in the agreement between ANN outputs and neural or behavioural data. Finally, the ability to replicate topographic elements of brain organization in ANNs allows for testing of such representational arrangements to better understand their origins and functional implications (reviewed elsewhere[47–49,51,183,185] for work in this direction).

Developmental agreement. Methods for finding behavioural and neural data agreement and in silico electrophysiology can be applied at different points during network training, from untrained to fully trained models, and the learning trajectories obtained can be compared with different stages in biological development for their level of agreement[186] (Fig. 4d). Although it is currently unclear which aspects of learning in ANNs are better seen as corresponding to learning during evolution and which are better seen as modelling learning during the lifetime of an organism[67], both can be addressed experimentally. For example, the age at which children learn different words can be predicted by the performance of ANNs trained on visual classification and captioning tasks, over and above the expected effect that more frequent words are learnt earlier[187].

In summary, neuroconnectionism has a large array of techniques for evaluating and contrasting the ability of different models to explain brain data, which are vital aspects of the research programme. Each model can be extensively tested for how well it maps onto brain processing across levels from single neurons to behaviour, which, arguably, no approach besides neuroconnectionism can claim.

Model interpretation

A strength of ANNs is that a single model can be fit to biological data across various levels, from the selectivity of a single neuron, through representational dynamics of neuronal populations, and onto behaviour, allowing for the study of these different scales in a single unified framework (Fig. 4). Yet, fitting data is not enough for true understanding, if the fitted model is an uninterpretable black box. Models also need to be 'transparent under analysis'[60] and enable researchers, via careful experimental manipulations and advanced analysis tools, to understand which aspects of the model are crucial to successfully account for brain data using the testing methods outlined earlier better than contrasting models. Next, we review such model interpretation tools that are available to neuroconnectionism and how, as a result of working with ANNs, researchers now have a whole new vocabulary for thinking about and describing brain computations underlying cognition.

Hypothesis testing via model contrasting. In the central research cycle of neuroconnectionism (Fig. 1), falsifiable belt hypotheses are evaluated by building, training and testing ANNs against neural and behavioural data. In line with standard scientific practice, hypotheses need to be tested against alternative hypotheses. Therefore, reporting how much variance a single ANN explains is not insightful (for example, even random models can explain some variance in neural recordings[188]). Rather, models must be contrasted to understand the relative impact of different model design choices. The underlying experimental logic is similar to more classic approaches in cognitive neuroscience that contrast different experimental conditions to understand which aspects better explain neural data. In the case of neuroconnectionism, hypotheses can be formed about the architecture of the ANN, the input data driving the network, its objective function and its learning rules[52,56]. By contrasting these hypotheses — instantiated as trained models — insights into the types of brain computations that lead to cognition are obtained. For example, by contrasting feedforward and

recurrent ANNs trained on the same data set, objective and learning rule, it was shown that the recurrent ANNs better match human neural dynamics[23] and behavioural reaction times[155].

Normative modelling. Normative modelling enables researchers to ask which task or which objective a system (for example, from neural populations, brain regions, all the way up to organisms) may be fulfilling — it thereby helps answer why a system is exhibiting the features it does[55,189,190]. This is done by optimizing a model to fulfil one (or many) pre-defined objective(s)[55], testing it against neural and behavioural data and contrasting it with models that are optimized towards fulfilling other objectives. Normative modelling dates back to earlier work on visual representations that underlie sparse coding[191], and has become used more widely with the rise of deep neural networks, for instance, by using task-trained neural networks to successfully predict ventral stream population responses[21,22,25,192,193]. As another example, it was shown that neurons in macaque face patches were better modelled by β-VAEs, a type of ANN that explicitly aims to disentangle sensory data into interpretable latent factors, compared with several other control ANNs without disentanglement as an objective to fulfil[50]. Thus, this normative modelling result suggests that face patches carry out computations with disentanglement as a goal. Recently, optimizing models to fulfil cross-modal objectives has also been explored. For example, recurrent ANNs were trained to map from input images to sentence-level linguistic embeddings of scene captions and better explained activity in large parts of visually driven brain areas when contrasted with more traditional category-trained networks[31]. Thus, such visuo-semantic transformations might offer a better conceptualization of the human visual system than object categorization.

New concepts for brain science. Working with and understanding ANNs provides a set of quantitative concepts that allow for fundamentally new ways of thinking about and describing the brain and its computations. Indeed, the vocabulary of neuroconnectionism revolves around technical terms (activation functions, layers, data set statistics, loss functions, learning rules), which are different from the terms previously used in neural or cognitive science. A recent example of a study revealing a new way of thinking about neural organization used a category-trained ANN to define a visual object space[46]. On the basis of this ANN-derived object space, the authors were not only able to predict neural selectivity in a previously uncharacterized 'no-man's land' of inferior temporal cortex but also to embed these novel findings into a unified picture of functional organization in that cortical region. Another study derived a new and mathematically tractable theory of semantic learning on the basis of the learning dynamics of ANNs, which mirrored many empirical phenomena of human semantic development[30]. Further examples proposed to explain adolescent changes in working memory by pruning in ANNs[194] and predicted the memorability of images in humans from response magnitude in ANN layers[195].

Neural control. As described in the 'Neural data agreement' section, successful predictions of neuronal firing rates from ANNs (for example, via using encoding models) open avenues for new experimental approaches, including precise control of neurons in biological brains. For example, by relying on the capacity of ANNs to predict neural activities for any given input image, researchers can vary stimuli to optimize their effect on firing rates of single neurons, groups of neurons or voxels[179–181]. The resulting images more strongly drive the neurons than any tested natural stimulus, aiding our understanding of their selectivity profiles beyond more classic experimental grounds.

Mechanistic understanding. Given the full access to all parts of ANNs, neuroconnectionist models can provide computational insights that are intractable without ANN simulations owing to the complexity of distributed neural codes and the difficulty of in vivo experiments. One example is 'model reduction', in which complex trained ANNs are reduced to simpler, more interpretable models. For instance, a study used an ANN (trained to mimic retinal ganglion cell outputs) to derive a simpler interpretable model of the retina, which was nevertheless able to explain several retinal phenomena[196]. Importantly, this approach combined the virtues of large-scale ANNs capable of dealing with natural stimuli — a crucial requirement for understanding the brain in naturalistic settings — with directly interpretable results. Mechanistic understanding can also be derived from in-depth experiments on models. For instance, a study that used an ANN to implement the feature similarity gain model of attention, in which attentional modulation is applied according to neural stimulus tuning, was able to provide an important test of this theory in naturalistic images[39]. The ANN model matched several experimentally observed neural and behavioural results and led to the novel mechanistic prediction that attention in the early visual cortex can be applied to optimize activities in the late visual system, rather than strictly according to tuning of early visual cortex neurons.

Formal theories of computation. As ANNs are heavily overparameterized and learn non-convex loss functions, precise mathematical tools are required to better understand the underlying computations and learning dynamics. Such tools are being developed in an emerging field known as 'deep mathematics'[197]. Insights from deep mathematics are of great importance for understanding complex neural processes, as the brain, too, is highly overparameterized[198]. The study of double descent in deep learning — the observation that increasing model complexity beyond the number of datapoints can nevertheless yield solutions that generalize well — is one such theoretical insight[199]. Another example is the study of neural tangent kernels, which indicated that training converges to a global optimum as networks get wider with more units per layer[200]. This finding helped answer the challenge that ANNs might converge to suboptimal local minima instead of a global optimum. Further theoretical work on ANNs has provided exact solutions to deep linear ANNs learning dynamics, explaining nonlinear phenomena during training such as plateaux and sudden dips in loss and proving that learning speed can occur independent of layer depth under certain conditions[142]. In addition, the geometry and symmetries of ANN loss landscapes — which represent the loss values in the weight space of the network — have been investigated in detail, revealing how permutation symmetries generate symmetry-induced critical points that help characterize the global minimum[201]. Advancing these lines of research by clarifying how they change under different architectures, data sets, objectives and learning rules will be an essential step towards understanding how ANNs compute, understanding why different ANNs map onto different aspects of brain processing and providing new formal tools to understand distributed computations in biological systems.

Explainable artificial intelligence in brain science. Substantial work in deep learning is devoted to network interpretability, aiming for better understanding of existing networks and for creating more interpretable ones (reviewed elsewhere[202]). Many of these deep

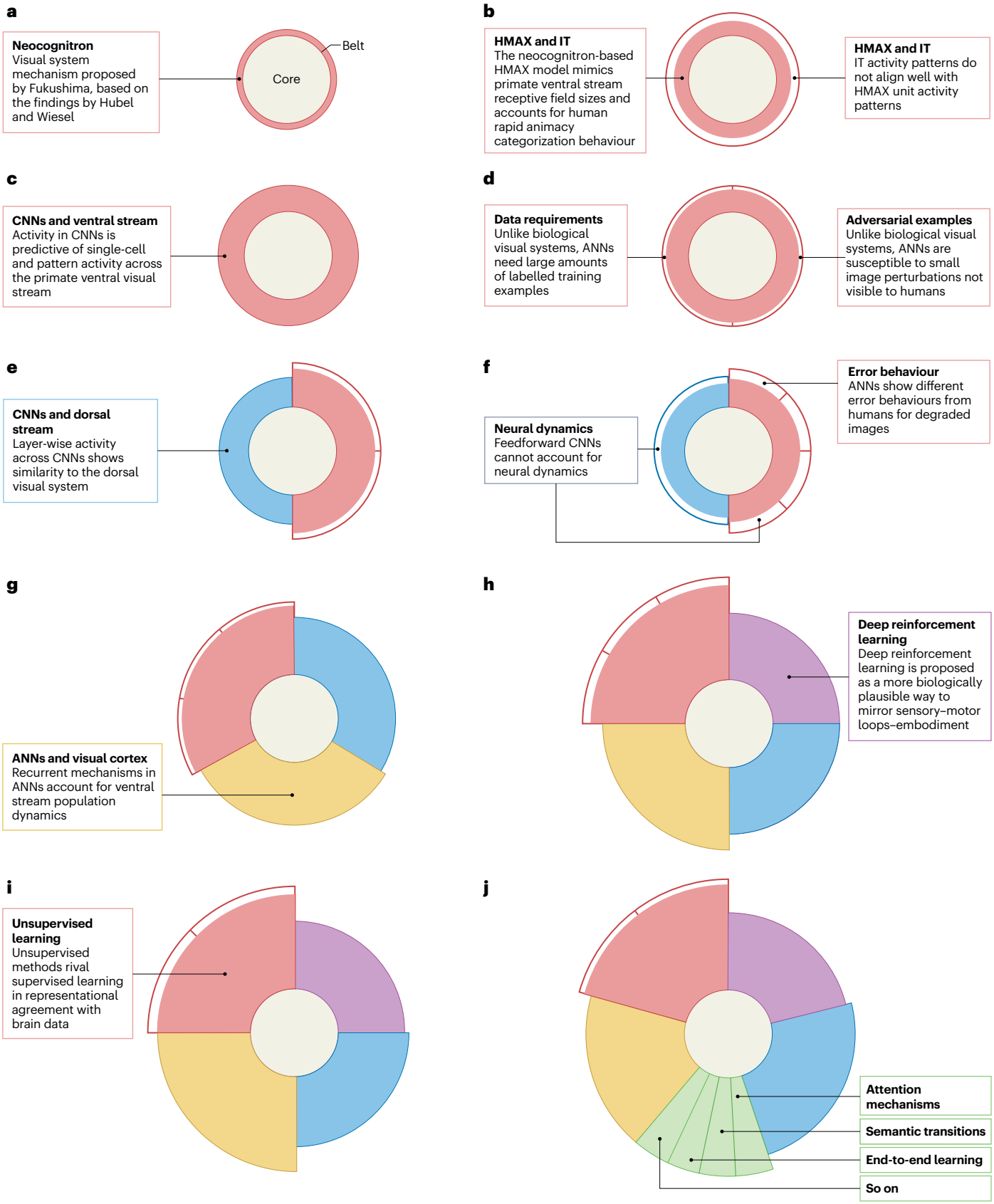

a
Neocognitron
Visual system mechanism proposed by Fukushima, based on the findings by Hubel and Wiesel
Core
Belt
b
HMAX and IT
The neocognitron-based HMAX model mimics primate ventral stream receptive field sizes and accounts for human rapid animacy categorization behaviour
HMAX and IT
IT activity patterns do not align well with HMAX unit activity patterns
c
CNNs and ventral stream
Activity in CNNs is predictive of single-cell and pattern activity across the primate ventral visual stream
d
Data requirements
Unlike biological visual systems, ANNs need large amounts of labelled training examples
Adversarial examples
Unlike biological visual systems, ANNs are susceptible to small image perturbations not visible to humans
e
CNNs and dorsal stream
Layer-wise activity across CNNs shows similarity to the dorsal visual system
f
Error behaviour
ANNs show different error behaviours from humans for degraded images
Neural dynamics
Feedforward CNNs cannot account for neural dynamics
g
ANNs and visual cortex
Recurrent mechanisms in ANNs account for ventral stream population dynamics
h
Deep reinforcement learning
Deep reinforcement learning is proposed as a more biologically plausible way to mirror sensory–motor loops–embodiment
i
Unsupervised learning
Unsupervised methods rival supervised learning in representational agreement with brain data
j
Attention mechanisms
Semantic transitions
End-to-end learning
So on

Fig. 5 | The historical progression of the neuroconnectionism belt in visual computational neuroscience is highly progressive. Empirical and analytical findings corroborating belt hypotheses are integrated into and strengthen the belt (increase or addition of filled areas), whereas disconfirmatory evidence (white bands that reduce the area of the belt hypothesis) poses challenges that need to be addressed. **a**, The neocognitron[20] was derived from seminal findings about simple and complex visual system cells by Hubel and Wiesel[296]. It learned and recognized increasingly abstract visual patterns through mechanisms that were similar to convolutions. **b**, HMAX[297], a more powerful model based on similar computations as in the earlier neocognitron, was investigated closely for its similarity to human behaviour and primate neural activity. HMAX was shown to match well to human psychophysical data on animacy detection well, thereby corroborating and strengthening this item in the belt but did not align well with broad activity patterns observed in IT, providing disconfirmatory evidence and weakening the belt item. **c**, Convolutional neural networks (CNNs) were successfully trained on large collections of naturalistic images. A series of studies showed that their layer activities match neural activity patterns along the primate ventral visual stream[21,22,25]. This was the first time that a single image-computable and functional object recognition network was able to match activity patterns across the ventral visual system. These findings resolved the objection to HMAX and strengthened the neuroconnectionist belt, which spawned a series of new neuroconnectionist studies. **d**, With their susceptibility to adversarial attacks[249–251] and the amounts of labelled training data they required, CNNs were shown to exhibit several important differences with biological vision. **e**, CNNs showed similar layer activities to the dorsal visual stream, adding further experimental evidence strengthening the belt[229,298]. **f**, Removing support from the belt, it was shown that the error behaviour during image alterations clearly diverges between humans and CNNs[156,252]. Furthermore, feedforward CNNs embodied too simple mechanisms to cover neural dynamic observations beyond coarse rate coding. **g**, The neural dynamics objection was resolved by the demonstration that dynamic transformations during visual processing can be captured if recurrence is added to artificial neural networks (ANNs)[23,90,218]. This resolved one of the challenges facing CNNs and thereby strengthened the neuroconnectionist belt. **h**, It was shown that activity across the dorsal visual stream during game playing matches activity in deep reinforcement learning networks[41], which implement a sensory–motor loop for the same game playing tasks. **i**, Newer developments demonstrated that unsupervised learning can rival supervised learning in representational agreement with brain data[121,225,227]. This finding solved the challenge that too many labelled examples were needed for training, further strengthening the belt. **j**, Future directions. Attention mechanisms, semantic objectives and end-to-end learning in which networks trained directly to match neural activity are recent developments in ANNs. Future experiments will reveal which brain processes are better modelled by incorporating these elements.

learning approaches can also be used in brain science[203]. Feature visualization methods investigate what ANNs detect by quantifying which parts of the input most strongly impact unit activities and network behaviour[204–206]. Feature attribution techniques dissect how ANNs process different inputs, by determining the contribution of different parts of the network (for example, different units, channels or layers)[207,208]. As another example, textual justification is an approach in which a model is trained to provide textual explanations of how it reached its decision, for example, by providing a text description of which visual features were most salient to the network[209]. This fast expanding set of interpretative tools for ANNs can help understand how they compute, helping to better shed light on the brain computations which they model.

Links to higher-level cognitive neuroscience and psychology theories. ANNs can be related to existing psychological and psychophysical ideas and might be crucial for unifying often disparate theories in these fields. As described earlier, ANNs can be tested on traditional psychophysical effects and the neuroconnectionist toolbox can be used to understand which model aspects are crucial to explain these effects. For example, convolutional neural networks (CNNs) failed to account for important global effects in visual crowding, a strong psychophysical phenomenon[159,210], indicating that an important computational feature is missing from CNNs. In psychology, these global crowding effects are often explained by 'perceptual grouping', an imprecise and highly abstract concept[211,212] that can be difficult to test experimentally, given that it cannot be implemented in a mechanistic model. By contrast, ANNs can be used to formulate precise explanations and mechanical models, which can be tested across many experiments. For instance, Capsule networks[213], a class of image-computable neuroconnectionist models that combine CNN feature extraction with a recurrent grouping and segmentation process in an ANN, solved the problems of traditional CNNs across varied psychophysical stimuli related to visual crowding[89,214,215]. Beyond the specific case of visual crowding, there are different psychological models for different paradigms (for example, a traditional psychological model tailored for visual search cannot be directly applied to a memory setting). A goal of neuroconnectionism is to find ANNs that unify these models, which is only possible because ANNs can process any image (or other modality) as input, thus going beyond models specific to a given psychological effect or a set of theoretical constructs in psychology.

The neuroconnectionist belt

We have now defined the core of the research programme and the neuroconnectionist tools that are used to design, train and evaluate ANN models across various levels of explanation to derive new understanding of the brain computations underlying cognition. Next, we discuss the current belt of the research programme, a set of auxiliary hypotheses which are tested and which evolve as new empirical data are integrated. Individual elements of the belt are important, but a more central aim, when taking a Lakatosian perspective, is an evaluation of longitudinal developments (both theoretical and empirical), which determine whether a research programme is progressive or degenerative (Fig. 2). New hypotheses can be derived and existing hypotheses can be corroborated, altered and rejected so that the belt of a research programme is subject to change. An individual belt hypothesis that is rejected does not refute the core assumptions upon which a research programme is built, but rather provides an important datapoint for future developments. According to this Lakatosian view, the overarching question becomes: How does the neuroconnectionism research programme fare in terms of productivity? Subsequently, we discuss whether neuroconnectionism generates new insights, and how well it addresses existing challenges. We then examine whether challenges that the research programme has not overcome are roadblocks rendering it degenerative or are signposts towards open questions improving the research programme that render it progressive.

The progressive evolution of the neuroconnectionist belt

The neuroconnectionist belt has considerably evolved in the past decade. By rapidly testing and expanding the number of belt hypotheses, the research programme now commonly uses ANNs with different architectures, training data sets, objectives and learning rules to test

the resulting models across various experimental settings for their alignment with brain and behavioural data.

One of the clearest examples of progressive evolution of neuroconnectionism has been the change in how vision is modelled with ANNs in recent years (Fig. 5). In the early 2010s, researchers focused a great deal of effort on comparing neural and behavioural data to what were, at that time, the state of the art in ANNs for vision, namely, deep CNNs trained on image classification tasks[19–22,24,25]. Over time, as researchers explored the successes and the failures of these models, the field has seen almost every component of these initial deep CNN models updated, leading to new models that better account for neurophysiological and psychological data. The earliest such update resulted from the recognition that the feedforward nature of standard CNNs was both obviously incongruent with real neural anatomy and functionally limiting[216]. Adding recurrence to networks improved matching both behavioural data and neural activity patterns that occurred with longer delays[23,38,89–93,95,165,217–220]. In addition, researchers explored ways to improve the training data sets beyond computer vision benchmarks, including a more ecologically relevant selection of object categories[119], video data[221], embodiment[222,223] and goal-directed eye movements[224]. Similarly, the use of supervised category training, which was always problematic from a neuroscience perspective as humans do not need millions of labelled examples to learn, was shown to be unnecessary: self-supervised techniques for training ANNs, for example, objectives that do not require training labels but rather bootstrap information from the data set itself lead to as good or better matches to neural representation and animal behaviour[121,225,226]. Moreover, self-supervised training on video data can account for the functional distinction between the dorsal and ventral pathways in the brain[227], and self-supervised training on lower-resolution inputs provides a better fit to electrophysiological data of mouse visual cortex[228]. Other loss functions have also been explored, and researchers have found that voxel or neuronal activity in the visual dorsal stream can be modelled by ANN units trained both by control-based optimization[41] and by self-motion-related loss functions[229].

A similar evolution has occurred in models of the hippocampal formation and related networks, in which initial architectures and loss functions have been replaced as the belt of the research programme evolved. Early ANN-based hippocampal models were often attractor networks[230,231], which captured many interesting aspects of the underlying circuitry supporting such networks. But, with time, these early models have evolved to incorporate additional architectural features and loss functions related to prediction and spatial integration, leading to improved matches with a host of experimental results[232–234]. Moreover, new ANN architectures such as transformers have been created with more sophisticated attention mechanisms, and research has demonstrated that transformers trained in a self-supervised manner can effectively capture the representations observed in language areas of the brain[32,33], and other brain circuits, such as the mnemonic circuits of the medial temporal lobes[235]. The success of transformers at explaining brain data in these areas leaves open the possibility for numerous other updates and explorations to existing models of vision and other senses, including the use of self-attention and the use of multimodal networks that combine linguistic inputs with vision or other sensory modalities[236,237].

Many developments in the belt are accompanied by methodological developments in the ways in which we train and test models. These include developments in quantifying the alignment of ANN and brain data[172,178,238,239] and training networks end-to-end directly to match neural data[23,90,106–113]. Further examples include work highlighting the important individual variability across network instances[182], explicitly pitting networks against each other using 'controversial stimuli'[240] and integrating hierarchical[76] and temporal[89,165] aspects of information processing into model comparisons. In addition, performing detailed psychophysical experiments[159,241,242], using metamers to compare humans and models[243], and developing interpretable low-rank recurrent networks[244] allow for greater interpretability.

Altogether, these rapid developments within neuroconnectionism demonstrate that the underlying research programme is highly progressive. Researchers have repeatedly updated their models, altering the architectures, objective functions and training data sets to arrive at a progressively better account of neural information processing, including novel ANN-driven insights and predictions. Although final answers remain distant and there are still important methodological issues[173,245–247], this progression illustrates how neuroconnectionist experiments, theory and methodology form a virtuous circle in which new empirical and theoretical results stimulate improved methodology. In turn, this allows for newly generated hypotheses to be tested, enabling the research programme to keep progressing.

Shortcomings as signposts, not roadblocks

Despite the advances in neuroconnectionism, skepticism of ANNs as models of brain function continues because differences with neural and behavioural data remain. In the following, we consider prominent examples of such differences, asking whether the underlying controversies are roadblocks that preclude progress (indicating neuroconnectionism is a degenerating research programme) or are better perceived as signposts that point towards promising directions for improvement (suggesting neuroconnectionism is a progressive research programme).

One of the main controversies surrounding ANNs in both cognitive science and artificial intelligence concerns differences in the behaviour of ANNs versus biological brains in various settings[59,62,248]. Although capable of impressive behaviours, current ANNs do not model all aspects of biological behaviour equally well: they sometimes generalize poorly[157], can be easily fooled[249–251] and behave differently from humans in many psychophysical settings[59,156,159,252,253]. Perhaps, the most famous of these differences is given by adversarial examples[249,250,254]: small perturbations to an image, invisible to the human eye, completely change an ANN classification of that image. This provokes fascinating new questions. First, how can ANNs be made more robust? One straightforward, yet technical, approach is to add adversarial images to the training data set[255]. This approach successfully increases the ANNs robustness, but it is a questionable strategy for models of human cognition. Second, can biological features of the visual system, such as V1-like receptive fields[256,257], be leveraged to increase robustness? Biologically inspired features have had some success increasing robustness, but work remains to fully understand their impact (reviewed elsewhere[258]). Third, where do adversarial examples come from? Perhaps, the most influential view is that they are features, not bugs[259]: they occur when ANNs latch onto real patterns in the data set that are highly predictive, but non-robust (because they are not present in other data sets) and undetectable by humans. Under this view, networks take a shortcut in learning[260] by harnessing specifics of a given data set to solve the task at hand, and the shortcut used is distinct from human cognitive processing. It is an open empirical question whether ANNs will become more human-like with richer, multimodal, more naturalistic data sets, other task settings and currently underexplored features such as embodiment. In summary, adversarial examples remain a fascinating tool for

understanding differences and similarities between visual information processing in brains and ANNs models[69,261–263].

Another important behavioural difference between humans and current ANNs is shape versus texture biases during object classification. Extensive psychological studies have shown that humans rely on global shape to classify objects, whereas, by contrast, current ANNs have a bias towards texture[156,252]. This difference during object classification clearly poses a challenge to current ANN models of human vision. As with adversarial examples, there are various approaches to address this shortcoming, including training ANNs on images with randomized textures[156] (but see[159] for an example in which texture randomization had negligible effects on the global processing abilities of CNNs), improving data augmentation[264], using a more biologically inspired first network layer[265] or training ANNs on substantially larger data sets[266]. This shape versus texture distinction was studied in the brain too, revealing that inferior temporal cortex might rely on a textural encoding[267], a hypothesis that strongly contrasts with the classic object-based view of the inferior temporal cortex function. Together, addressing this challenge illustrates how working with ANNs has shaped research questions about the brain and has likewise influenced experimental designs.

A further controversy around the computations of ANNs is focused on symbolic manipulations in ANNs versus biological brains. It has been argued that current ANNs, unlike humans, do not compute with symbols because they do not have systematicity, dynamic variable binding, role-filler independence or appropriately structured representations[268–271]. In addition, owing to the human capacity to understand and create novel combinations of concepts[272,273], ANNs are thought to run into tractability or scalability issues[268,270]. Yet, even ardent defenders of symbolism agree that ANNs should, in principle, be able to implement the desired symbolic systems[268,270,274], as such a capacity is an immediate result of the universal approximation theorems that hold for multilayer and recurrent neural networks[83,275]. Still, the issue of symbolism points to potentially fundamental computational differences between current ANNs and humans that neuroconnectionism needs to address. Indeed, the field of neurosymbolic computations tackles this open problem. One approach is centred around the idea that symbolic computations could arise from training in complex social and cultural environments[276,277]. If true, the question becomes: how are data sets that favour the emergence of symbols developed? Improving objective functions is also an active area of this field. By training ANNs on new objectives, for instance, those that are as demanding as the goals humans are able to attain, the models might be incentivized to develop symbolic reasoning. For example, recent large-language models use complex semantic objective functions, including interactive conversations with humans[278], which have helped bring models ever closer to human reasoning (with limitations such as fact hallucination and reasoning errors[279,280]). There are also architectural approaches that try to endow ANNs with inductive biases for symbolic representations. For instance, some approaches develop architectures that allow for less distributed, more disentangled representations, in which different features of the same object are represented by different population activities[213,281–284]. The idea here is that less distributed, disentangled representations can be combined more easily, which could cause the network to develop the type of compositionality that is required for generalized symbolic reasoning. Other architectural approaches go a step further and try to build ANNs with mechanisms that should directly enable these architectures to perform symbolic computations (for example, tensor products, pointers or explicit read–write addressable memory)[285–291].

In sum, these example challenges presented by adversarial examples, shape versus texture distinctions and symbolic computations are leading to new avenues of neuroconnectionism research, indicating that they are better viewed as signposts fostering progress within the research programme rather than roadblocks implying its degeneration.

Conclusions

In less than a decade, ANN modelling has gone from being fringe to being a more central research tool in many parts of cognitive neuroscience, including vision, audition, motor control, language and higher-level cognitive tasks. The underlying neuroconnectionism research programme has met striking successes with many researchers sharing the excitement about this new framework. However, the use of ANNs for brain science also faces challenges, and discussions thereof often escalate into binary debates on the usefulness of ANNs. We argued that this binary view is unhelpful and that, instead, a more longitudinal view in line with Lakatosian research programmes is more adequate, comprehensively describing neuroconnectionism, its philosophical underpinnings, scientific rationale, experimental tools and methods for understanding.

Traditional modelling approaches based on simple human-interpretable concepts cannot bridge levels of explanation from single neurons to behaviour while performing complex cognitive tasks that require rich domain knowledge and sensory grounding. Hence, more complex models are needed to complement simpler approaches. However, complex models come with their own challenges, including computational tractability and interpretability. Until recently, these challenges could not be met, but ANNs now offer a promising framework to overcome them. Indeed, ANNs are high-dimensional enough to carry out cognitive tasks, encode domain knowledge and be grounded in sensory input while still being interpretable enough to yield insights into brain computations. Hence, we argued that ANNs reside in a Goldilocks zone of biological abstraction: abstract enough to be computationally tractable and applied to complex cognitive tasks, whereas still being close enough to biology to incrementally test which biological detail is needed in a hypothesis-driven manner. This is possible owing to an extensive toolbox that has been developed allowing to create, train, test and understand ANNs.

The framework of progressive versus degenerating Lakatosian research programmes is well suited to evaluate the promise of this intricate research programme. Of central importance is the view that neuroconnectionism is not a single theory or hypothesis. Instead, it is a research programme composed of many different auxiliary hypotheses and research directions in the belt, each sharing the same core of desiderata, fixed background assumptions and distinctive features. On the basis of this view, we showed that current challenges represent important signposts that aid further progress, rather than roadblocks. Indeed, many such challenges have already sparked vibrant new research directions. Together with the growing body of studies demonstrating good agreement between ANNs and brain data, these observations illustrated that the neuroconnectionism research programme is highly progressive.

Although we have here focused on the progressive nature of the research programme, we do not mean to imply that the field is anywhere near to successfully explaining cognition or its underlying brain computations. Neuroconnectionism is in its infancy, and hence knowing where it fails is equally important as knowing where it works. In addition to addressing the challenges discussed here, future work will need to incorporate many currently missing aspects of lower and higher

cognition. These future directions include (i) multitask networks that can explain the human ability to perform multiple, sometimes highly abstract tasks based on sensory evidence[292], (ii) a focus on embodiment rather than treating networks as artificial brains in vats[222,228,229,293], (iii) data-efficient and continuous learning using biologically realistic learning rules and inductive biases, (iv) an answer to how symbolic reasoning is implemented in neuroconnectionist models, (v) better methods for unsupervised, multimodal learning, (vi) better modelling of cognitive development, (vii) the integration of multiple memory systems, (viii) models that learn in social contexts, (ix) methods taking into account that ANNs learn tabula rasa, whereas humans come to experiments with strong priors (acquired both through evolution and through life experiences)[294] and (x) creating models with more robust inference. The fact that all these aspects are currently missing but can in principle be implemented in ANNs perfectly illustrates both the current limitations and strong potential of the neuroconnectionist programme. Rather than feeling threatened by the possibility of a new connectionist winter, the neuroconnectionist community should therefore continue to welcome criticism and limitations as they point the way towards new insights. Critics of neuroconnectionism, on the contrary, should not regard every shortcoming of the current set of networks as a failure of the entire research programme. In line with Lakatosian philosophy[295], time will tell whether neuroconnectionism can deliver on its promises to explain the emergence of cognitive phenomena, behaviour and neural data from bio-inspired, yet simple distributed coding principles. For now, it remains a highly progressive and therefore exciting research programme that welcomes critical signposts to guide the way.

Published online: 30 May 2023

References

1. Churchland, P. S. & Sejnowski, T. J. Blending computational and experimental neuroscience. *Nat. Rev. Neurosci.* **17**, 667–668 (2016).
2. Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D. Neuroscience needs behaviour: correcting a reductionist bias. *Neuron* **93**, 480–490 (2017).
3. Kanwisher, N. & Yovel, G. The fusiform face area: a cortical region specialized for the perception of faces. *Philos. Trans. R. Soc. B Biol. Sci.* **361**, 2109–2128 (2006).
4. Sergent, J., Ohta, S. & Macdonald, B. Functional neuroanatomy of face and object processing: a positron emission tomography study. *Brain* **115**, 15–36 (1992).
5. Tong, F., Nakayama, K., Vaughan, J. T. & Kanwisher, N. Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron* **21**, 753–759 (1998).
6. Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B. & Tootell, R. B. Faces and objects in macaque cerebral cortex. *Nat. Neurosci.* **6**, 989–995 (2003).
7. Rust, N. C. & Movshon, J. A. In praise of artifice. *Nat. Neurosci.* **8**, 1647–1650 (2005).
8. Vinken, K., Konkle, T. & Livingstone, M. The neural code for 'face cells' is not face specific. Preprint at *bioRxiv* https://doi.org/10.1101/2022.03.06.483186 (2022).
9. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133 (1943).
10. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
11. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).
12. Schrimpf, M. et al. Brain-score: which artificial neural network for object recognition is most brain-like? Preprint at *bioRxiv* https://doi.org/10.1101/407007 (2020).
13. Cichy, R. M. et al. The Algonauts Project: a platform for communication between the sciences of biological and artificial intelligence. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1905.05675 (2019).
14. Allen, E. J. et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* **25**, 116–126 (2022).
15. Willeke, K. F. et al. The sensorium competition on predicting large-scale mouse primary visual cortex activity. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2206.08666 (2022).
16. RichardWebster, B., DiFalco, A., Caldesi, E. & Scheirer, W. J. Perceptual-score: a psychophysical measure for assessing the biological plausibility of visual recognition models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2210.08632 (2022).
17. Schlangen, D. Targeting the benchmark: on methodology in current natural language processing research. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2007.04792 (2020).
18. Rumelhart, D. E., McClelland, J. L. & Group, P. R. *Parallel Distributed Processing* Vol. 1 (IEEE, 1988).
19. Cichy, R. M., 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).
20. Fukushima, K. & Miyake, S. Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition. in *Competition and Cooperation in Neural Nets* 267–285 (Springer, 1982).
21. Guclu, U. & van Gerven, M. A. J. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).
22. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
23. Kietzmann, T. C. et al. Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl Acad. Sci. USA* **116**, 21854–21863 (2019).
24. Seeliger, K. et al. Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage* **180**, 253–266 (2018).
25. Yamins, D. L. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
26. Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V. & McDermott, J. H. A task-optimized neural network replicates human auditory behaviour, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* **98**, 630–644.e16 (2018).
27. Saddler, M. R., Gonzalez, R. & McDermott, J. H. Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nat. Commun.* **12**, 7278 (2021).
28. Cadena, S. A. et al. Diverse task-driven modeling of macaque V4 reveals functional specialization towards semantic tasks. Preprint at *bioRxiv* https://doi.org/10.1101/2022.05.18.492503 (2022).
29. Jackson, R. L., Rogers, T. T. & Lambon Ralph, M. A. Reverse-engineering the cortical architecture for controlled semantic cognition. *Nat. Hum. Behav.* **5**, 774–786 (2021).
30. Saxe, A. M., McClelland, J. L. & Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proc. Natl Acad. Sci. USA* **116**, 11537–11546 (2019).
31. Doerig, A. et al. Semantic scene descriptions as an objective of human vision. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2209.11737 (2022).
32. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5**, 134 (2022).
33. Schrimpf, M. et al. The neural architecture of language: integrative modeling converges on predictive processing. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.2105646118 (2021).
34. Hannagan, T., Agrawal, A., Cohen, L. & Dehaene, S. Emergence of a compositional neural code for written words: recycling of a convolutional neural network for reading. *Proc. Natl Acad. Sci. USA* **118**, e2104779118 (2021).
35. Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J. & Kurth-Nelson, Z. Deep reinforcement learning and its neuroscientific implications. *Neuron* **107**, 603–616 (2020).
36. Dabney, W. et al. A distributional code for value in dopamine-based reinforcement learning. *Nature* **577**, 671–675 (2020).
37. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
38. Quax, S. & van Gerven, M. Emergent mechanisms of evidence integration in recurrent neural networks. *PLoS ONE* **13**, e0205676 (2018).
39. Lindsay, G. W. & Miller, K. D. How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife* **7**, e38105 (2018).
40. Orhan, A. E. & Ma, W. J. A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nat. Neurosci.* **22**, 275–283 (2019).
41. Cross, L., Cockburn, J., Yue, Y. & O'Doherty, J. P. Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments. *Neuron* **109**, 724–738.e7 (2021).
42. Feulner, B. et al. Small, correlated changes in synaptic connectivity may facilitate rapid motor learning. *Nat. Commun.* **13**, 5163 (2022).
43. Merel, J., Botvinick, M. & Wayne, G. Hierarchical motor control in mammals and machines. *Nat. Commun.* **10**, 5489 (2019).
44. Michaels, J. A., Schaffelhofer, S., Agudelo-Toro, A. & Scherberger, H. A goal-driven modular neural network predicts parietofrontal neural dynamics during grasping. *Proc. Natl Acad. Sci. USA* **117**, 32124–32135 (2020).
45. Sussillo, D., Churchland, M. M., Kaufman, M. T. & Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* **18**, 1025–1033 (2015).
46. Bao, P., She, L., McGill, M. & Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
47. Blauch, N. M., Behrmann, M. & Plaut, D. C. A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proc. Natl Acad. Sci. USA* **119**, e2112566119 (2022).
48. Dobs, K., Martinez, J., Kell, A. J. E. & Kanwisher, N. Brain-like functional specialization emerges spontaneously in deep neural networks. *Sci. Adv.* **8**, eabl8913 (2022).
49. Doerig, A., Krahmer, B. & Kietzmann, T. Emergence of topographic organization in a non-convolutional deep neural network (Neuromatch 40). *Perception* **51**, 74–75 (2022).
50. Higgins, I. et al. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nat. Commun.* **12**, 6456(2021).
51. Lee, H. et al. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. Preprint at *bioRxiv* https://doi.org/10.1101/2020.07.09.185116 (2020).

52. Kietzmann, T. C., McClure, P. & Kriegeskorte, N. Deep neural networks in computational neuroscience. *Neuroscience* https://doi.org/10.1093/acrefore/9780190264086.013.46 (2019).
53. Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015).
54. Lindsay, G. W. Convolutional neural networks as a model of the visual system: past, present, and future. *J. Cogn. Neurosci.* **33**, 2017–2031 (2021).
55. Marblestone, A. H., Wayne, G. & Kording, K. P. Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.* **10**, 94 (2016).
56. Richards, B. A. et al. A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).
57. Saxe, A., Nelli, S. & Summerfield, C. If deep learning is the answer, what is the question? *Nat. Rev. Neurosci.* **22**, 55–67 (2020).
58. Van Gerven, M. Computational foundations of natural intelligence. *Front. Comput. Neurosci.* **11**, 112 (2017).
59. Bowers, J. S. et al. Deep problems with neural network models of human vision. *Behav. Brain Sci.* https://doi.org/10.1017/S0140525X22002813 (2022).
60. Leek, E. C., Leonardis, A. & Heinke, D. Deep neural networks and image classification in biological vision. *Vis. Res.* **197**, 108058 (2022).
61. Marcus, G. Deep learning: a critical appraisal. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1801.00631 (2018).
62. Serre, T. Deep learning: the good, the bad, and the ugly. *Annu. Rev. Vis. Sci.* **5**, 399–426 (2019).
63. Cao, R. & Yamins, D. Explanatory models in neuroscience: part 1 — taking mechanistic abstraction seriously. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2104.01490 (2021).
64. Cichy, R. M. & Kaiser, D. Deep neural networks as scientific models. *Trends Cogn. Sci.* **23**, 305–317 (2019).
65. Storrs, K. R. & Kriegeskorte, N. Deep learning for cognitive neuroscience. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1903.01458 (2019).
66. Barrett, D. G., Morcos, A. S. & Macke, J. H. Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Curr. Opin. Neurobiol.* **55**, 55–64 (2019).
67. Zador, A. M. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat. Commun.* **10**, 3770 (2019).
68. Yang, G. R. & Wang, X.-J. Artificial neural networks for neuroscientists: a primer. *Neuron* **107**, 1048–1070 (2020).
69. Wichmann, F. A. & Geirhos, R. Are deep neural networks adequate behavioural models of human visual perception? *Annu. Rev. Vis. Sci.* https://doi.org/10.1146/annurev-vision-120522-031739 (2023).
70. Pulvermüller, F., Tomasello, R., Henningsen-Schomers, M. R. & Wennekers, T. Biological constraints on neural network models of cognitive function. *Nat. Rev. Neurosci.* **22**, 488–502 (2021).
71. Lakatos, I. Falsification and the methodology of scientific research programmes. in *Can Theories Be Refuted?* 205–259 (Springer, 1976).
72. Anderson, J. R., Matessa, M. & Lebiere, C. ACT-R: a theory of higher level cognition and its relation to visual attention. *Hum. Comput. Interact.* **12**, 439–462 (1997).
73. Wittgenstein, L. *Philosophical Investigations* (John Wiley & Sons, 2009).
74. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. in *Advances in Neural Information Processing Systems* 1097–1105 (ACM, 2012).
75. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1409.1556 (2014).
76. Nonaka, S., Majima, K., Aoki, S. C. & Kamitani, Y. Brain hierarchy score: which deep neural networks are hierarchically brain-like? *iScience* **24**, 103013 (2021).
77. Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P. & De Lange, F. P. A. hierarchy of linguistic predictions during natural language comprehension. *Proc. Natl Acad. Sci. USA* **119**, e2201968119 (2022).
78. Ponce, C. R. et al. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell* **177**, 999–1009.e10 (2019).
79. Tuli, S., Dasgupta, I., Grant, E. & Griffiths, T. L. Are convolutional neural networks or transformers more like human vision? Preprint at *arXiv* https://doi.org/10.48550/arXiv.2105.07197 (2021).
80. Markram, H. The human brain project. *Sci. Am.* **306**, 50–55 (2012).
81. Nandi, A. et al. Single-neuron models linking electrophysiology, morphology, and transcriptomics across cortical cell types. *Cell Rep.* **40**, 111176 (2022).
82. Wolfram, S. Cellular automata as models of complexity. *Nature* **311**, 419–424 (1984).
83. Siegelmann, H. T. & Sontag, E. D. On the computational power of neural nets. *J. Comput. Syst. Sci.* **50**, 132–150 (1995).
84. Ali, A., Ahmad, N., de Groot, E., van Gerven, M. A. J. & Kietzmann, T. C. Predictive coding is a consequence of energy efficiency in recurrent neural networks. *Patterns* **3**, 100639 (2022).
85. Jaeger, H. The 'echo state' approach to analysing and training recurrent neural networks — with an erratum note. *Bonn. Ger. Ger. Natl Res. Cent. Inf. Technol. GMD Tech. Rep.* **148**, 13 (2001).
86. Maass, W., Natschläger, T. & Markram, H. Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput.* **14**, 2531–2560 (2002).
87. LeCun, Y. et al. Handwritten digit recognition with a back-propagation network. in *Advances in Neural Information Processing Systems* 396–404 (NIPS, 1990).
88. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
89. Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M. & Herzog, M. H. Capsule networks as recurrent models of grouping and segmentation. *PLoS Comput. Biol.* **16**, e1008017 (2020).
90. Güçlü, U. & Van Gerven, M. A. Modeling the dynamics of human brain activity with recurrent neural networks. *Front. Comput. Neurosci.* **11**, 7 (2017).
91. Kar, K. & DiCarlo, J. J. Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron* **109**, 164–176.e5 (2021).
92. Lindsay, G. W., Mrsic-Flogel, T. D. & Sahani, M. Bio-inspired neural networks implement different recurrent visual processing strategies than task-trained ones do. Preprint at *bioRxiv* https://doi.org/10.1101/2022.03.07.483196 (2022).
93. Linsley, D., Kim, J. & Serre, T. Sample-efficient image segmentation through recurrence. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1811.11356 (2018).
94. Nayebi, A. et al. Goal-driven recurrent neural network models of the ventral visual stream. Preprint at *bioRxiv* https://doi.org/10.1101/2021.02.17.431717 (2021).
95. Thorat, S., Aldegheri, G. & Kietzmann, T. C. Category-orthogonal object features guide information processing in recurrent neural networks trained for object categorization. *Preprint at arXiv* https://doi.org/10.48550/arXiv.2111.07898 (2021).
96. Bertalmío, M. et al. Evidence for the intrinsically nonlinear nature of receptive fields in vision. *Sci. Rep.* **10**, 16277 (2020).
97. Quax, S. C., D'Asaro, M. & van Gerven, M. A. Adaptive time scales in recurrent neural networks. *Sci. Rep.* **10**, 11360 (2020).
98. Voelker, A., Kajić, I. & Eliasmith, C. Legendre memory units: continuous-time representation in recurrent neural networks. in *Advances in Neural Information Processing Systems* Vol. 32 (NeurIPS, 2019).
99. Bohte, S. M. The evidence for neural information processing with precise spike-times: a survey. *Nat. Comput.* **3**, 195–206 (2004).
100. Gerstner, W. & Kistler, W. M. *Spiking Neuron Models: Single Neurons, Populations, Plasticity* (Cambridge Univ. Press, 2002).
101. Sörensen, L. K., Zambrano, D., Slagter, H. A., Bohté, S. M. & Scholte, H. S. Leveraging spiking deep neural networks to understand the neural mechanisms underlying selective attention. *J. Cogn. Neurosci.* **34**, 655–674 (2022).
102. Zenke, F. & Ganguli, S. Superspike: supervised learning in multilayer spiking neural networks. *Neural Comput.* **30**, 1514–1541 (2018).
103. Stimberg, M., Brette, R. & Goodman, D. F. Brian 2, an intuitive and efficient neural simulator. *eLife* **8**, e47314 (2019).
104. Guerguiev, J., Lillicrap, T. P. & Richards, B. A. Towards deep learning with segregated dendrites. *eLife* **6**, e22901 (2017).
105. Sacramento, J., Ponte Costa, R., Bengio, Y. & Senn, W. Dendritic cortical microcircuits approximate the backpropagation algorithm. in *Advances in Neural Information Processing Systems* Vol. 31 (NeurIPS, 2018).
106. Antolík, J., Hofer, S. B., Bednar, J. A. & Mrsic-Flogel, T. D. Model constrained by visual hierarchy improves prediction of neural responses to natural scenes. *PLoS Comput. Biol.* **12**, e1004927 (2016).
107. Cadena, S. A. et al. Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Comput. Biol.* **15**, e1006897 (2019).
108. Ecker, A. S. et al. A rotation-equivariant convolutional neural network model of primary visual cortex. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1809.10504 (2018).
109. Kindel, W. F., Christensen, E. D. & Zylberberg, J. Using deep learning to probe the neural code for images in primary visual cortex. *J. Vis.* **19**, 29–29 (2019).
110. Klindt, D., Ecker, A. S., Euler, T. & Bethge, M. Neural system identification for large populations separating 'what' and 'where'. in *Advances in Neural Information Processing Systems* Vol. 30 (NIPS, 2017).
111. Seeliger, K. et al. End-to-end neural system identification with neural information flow. *PLoS Comput. Biol.* **17**, e1008558 (2021).
112. St-Yves, G. & Naselaris, T. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *NeuroImage* **180**, 188–202 (2018).
113. Tripp, B. Approximating the architecture of visual cortex in a convolutional network. *Neural Comput.* **31**, 1551–1591 (2019).
114. Bellec, P. & Boyle, J. Bridging the gap between perception and action: the case for neuroimaging. Preprint at *PsyarXiv* https://doi.org/10.31234/osf.io/3epws (2019).
115. Hebart, M. N. et al. THINGS: a database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS ONE* **14**, e0223792 (2019).
116. Naselaris, T., Allen, E. & Kay, K. Extensive sampling for complete models of individual brains. *Curr. Opin. Behav. Sci.* **40**, 45–51 (2021).
117. Seeliger, K., Sommers, R. P., Güçlü, U., Bosch, S. E. & Van Gerven, M. A. J. A large single-participant fMRI dataset for probing brain responses to naturalistic stimuli in space and time. Preprint at *bioRxiv* https://doi.org/10.1101/687681 (2019).
118. Siegle, J. H. et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature* **592**, 86–92 (2021).
119. Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N. & Kietzmann, T. C. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proc. Natl Acad. Sci. USA* **118**, e2011417118 (2021).
120. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. in *International Conference on Machine Learning* 1597–1607 (PMLR, 2020).

121. Konkle, T. & Alvarez, G. A. A self-supervised domain-general learning framework for human ventral stream representation. Preprint at *Nat. Commun.* **13**, 491 (2020).
122. Choksi, B. et al. Predify: augmenting deep neural networks with brain-inspired predictive coding dynamics. *Adv. Neural Inf. Process. Syst.* **34**, 14069–14083 (2021).
123. Lotter, W., Kreiman, G. & Cox, D. A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nat. Mach. Intell.* **2**, 210–219 (2020).
124. Soulos, P. & Isik, L. Disentangled face representations in deep generative models and the human brain. in *NeurIPS 2020 Workshop SVRHM* (NeurIPS, 2020).
125. Storrs, K. R., Anderson, B. L. & Fleming, R. W. Unsupervised learning predicts human perception and misperception of gloss. *Nat. Hum. Behav.* **5**, 1402–1417 (2021).
126. Franzius, M., Sprekeler, H. & Wiskott, L. Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Comput. Biol.* **3**, e166 (2007).
127. Franzius, M., Wilbert, N. & Wiskott, L. Invariant object recognition with slow feature analysis. in *International Conference on Artificial Neural Networks* 961–970 (Springer, 2008).
128. Kayser, C., Einhäuser, W., Dümmer, O., König, P. & Körding, K. Extracting slow subspaces from natural videos leads to complex cells. in *Artificial Neural Networks — ICANN 2001* Vol. 2130 (eds Dorffner, G., Bischof, H. & Hornik, K.) 1075–1080 (Springer, 2001).
129. Wiskott, L. & Sejnowski, T. J. Slow feature analysis: unsupervised learning of invariances. *Neural Comput.* **14**, 715–770 (2002).
130. Wyss, R., König, P. & Verschure, P. F. J. A model of the ventral visual system based on temporal stability and local memory. *PLoS Biol.* **4**, e120 (2006).
131. Lindsay, G. W., Merel, J., Mrsic-Flogel, T. & Sahani, M. Divergent representations of ethological visual inputs emerge from supervised, unsupervised, and reinforcement learning. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2112.02027 (2021).
132. Dwivedi, K., Bonner, M. F., Cichy, R. M. & Roig, G. Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS Comput. Biol.* **17**, e1009267 (2021).
133. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
134. Ahmad, N., Schrader, E. & van Gerven, M. Constrained parameter inference as a principle for learning. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2203.13203 (2022).
135. Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J. & Hinton, G. Backpropagation and the brain. *Nat. Rev. Neurosci.* **21**, 335–346 (2020).
136. Lillicrap, T. P., Cownden, D., Tweed, D. B. & Akerman, C. J. Random synaptic feedback weights support error backpropagation for deep learning. *Nat. Commun.* **7**, 13276 (2016).
137. Pozzi, I., Bohte, S. & Roelfsema, P. Attention-gated brain propagation: how the brain can implement reward-based error backpropagation. *Adv. Neural Inf. Process. Syst.* **33**, 2516–2526 (2020).
138. Richards, B. A. & Lillicrap, T. P. Dendritic solutions to the credit assignment problem. *Curr. Opin. Neurobiol.* **54**, 28–36 (2019).
139. Hebb, D. O. *The Organization of Behaviour: A Neuropsychological Theory* (Psychology Press, 2005).
140. Rao, R. P. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79 (1999).
141. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59–69 (1982).
142. Saxe, A. M., McClelland, J. L. & Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1312.6120 (2013).
143. Benjamin, A. S., Zhang, L.-Q., Qiu, C., Stocker, A. & Kording, K. P. Efficient neural codes naturally emerge through gradient descent learning. *Nat. Commun.* **13**, 7972 (2022).
144. Munakata, Y. & Pfaffly, J. Hebbian learning and development. *Dev. Sci.* **7**, 141–148 (2004).
145. Berrios, W. & Deza, A. Joint rotational invariance and adversarial training of a dual-stream transformer yields state of the art brain-score for area V4. Preprint at https://doi.org/10.48550/arXiv.2203.06649 (2022).
146. St-Yves, G., Allen, E. J., Wu, Y., Kay, K. & Naselaris, T. Brain-optimized neural networks learn non-hierarchical models of representation in human visual cortex. Preprint at *bioRxiv* https://doi.org/10.1101/2022.01.21.477293 (2022).
147. Hasenstaub, A., Otte, S., Callaway, E. & Sejnowski, T. J. Metabolic cost as a unifying principle governing neuronal biophysics. *Proc. Natl Acad. Sci. USA* **107**, 12329–12334 (2010).
148. Stone, J. V. *Principles of Neural Information Theory: Computational Neuroscience and Metabolic Efficiency* (*Tutorial Introductions*) (Tutorial Introductions, 2018).
149. Wang, Z., Wei, X.-X., Stocker, A. A. & Lee, D. D. Efficient neural codes under metabolic constraints. in *Advances in Neural Information Processing Systems* Vol. 29 (NIPS, 2016).
150. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
151. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2010.11929 (2020).
152. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
153. Mnih, V. et al. Playing Atari with deep reinforcement learning. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1312.5602 (2013).
154. Vinyals, O. et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575**, 350–354 (2019).
155. Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I. & Kriegeskorte, N. Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS Comput. Biol.* **16**, e1008215 (2020).
156. Geirhos, R. et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. in *International Conference on Learning Representations* (ICLR, 2018).
157. Geirhos, R. et al. Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems* Vol. 31 (NIPS, 2018).
158. Singer, J. J., Seeliger, K., Kietzmann, T. C. & Hebart, M. N. From photos to sketches-how humans and deep neural networks process objects across different levels of visual abstraction. *J. Vis.* **22**, 4 (2022).
159. Doerig, A., Bornet, A., Choung, O. H. & Herzog, M. H. Crowding reveals fundamental differences in local vs. global processing in humans and machines. *Vis. Res.* **167**, 39–45 (2020).
160. Funke, C. M. et al. Comparing the ability of humans and DNNs to recognise closed contours in cluttered images. in *18th Annual Meeting of the Vision Sciences Society* (*VSS 2018*) 213 (VSS, 2018).
161. Jacob, G., Pramod, R. T., Katti, H. & Arun, S. P. Qualitative similarities and differences in visual object representations between brains and deep networks. *Nat. Commun.* **12**, 1872 (2021).
162. Kim, J., Linsley, D., Thakkar, K. & Serre, T. Disentangling neural mechanisms for perceptual grouping. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1906.01558 (2019).
163. Loke, J. et al. A critical test of deep convolutional neural networks' ability to capture recurrent processing in the brain using visual masking. *J. Cogn. Neurosci.* **34**, 2390–2405 (2022).
164. RichardWebster, B., Anthony, S. & Scheirer, W. Psyphy: a psychophysics driven evaluation framework for visual recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 41 (IEEE, 2018).
165. Sörensen, L. K., Bohté, S. M., De Jong, D., Slagter, H. A. & Scholte, H. S. Mechanisms of human dynamic object recognition revealed by sequential deep neural networks. Preprint at *bioRxiv* https://doi.org/10.1101/2022.04.06.487259 (2022).
166. Firestone, C. Performance vs. competence in human–machine comparisons. *Proc. Natl Acad. Sci. USA* **117**, 26562–26571 (2020).
167. Lonnqvist, B., Bornet, A., Doerig, A. & Herzog, M. H. A comparative biology approach to DNN modeling of vision: a focus on differences, not similarities. *J. Vis.* **21**, 17–17 (2021).
168. Ma, W. J. & Peters, B. A neural network walks into a lab: towards using deep nets as models for human behaviour. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2005.02181 (2020).
169. Neri, P. Deep networks may capture biological behaviour for shallow, but not deep, empirical characterizations. *Neural Netw.* **152**, 244–266 (2022).
170. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
171. Kriegeskorte, N. & Wei, X.-X. Neural tuning and representational geometry. *Nat. Rev. Neurosci.* **22**, 703–718 (2021).
172. Kaniuth, P. & Hebart, M. N. Feature-reweighted representational similarity analysis: a method for improving the fit between computational models, brains, and behaviour. *NeuroImage* **257**, 119294 (2022).
173. Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J. & Kriegeskorte, N. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *J. Cogn. Neurosci.* **33**, 2044–2064 (2021).
174. Kornblith, S., Norouzi, M., Lee, H. & Hinton, G. Similarity of neural network representations revisited. in *International Conference on Machine Learning* 3519–3529 (PMLR, 2019).
175. Kriegeskorte, N. & Diedrichsen, J. Peeling the onion of brain representations. *Annu. Rev. Neurosci.* **42**, 407–432 (2019).
176. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *NeuroImage* **56**, 400–410 (2011).
177. van Gerven, M. A. J. A primer on encoding models in sensory neuroscience. *J. Math. Psychol.* **76**, 172–183 (2017).
178. Sexton, N. J. & Love, B. C. Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Sci. Adv.* **8**, eabm2219 (2022).
179. Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deep image synthesis. *Science* **364**, aav9436 (2019).
180. Gu, Z. et al. NeuroGen: activation optimized image synthesis for discovery neuroscience. *NeuroImage* **247**, 118812 (2022).
181. Ratan Murty, N. A., Bashivan, P., Abate, A., DiCarlo, J. J. & Kanwisher, N. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nat. Commun.* **12**, 5540 (2021).
182. Mehrer, J., Spoerer, C. J., Kriegeskorte, N. & Kietzmann, T. C. Individual differences among deep neural network models. *Nat. Commun.* **11**, 5725 (2020).
183. Doshi, F. R. & Konkle, T. Visual object topographic motifs emerge from self-organization of a unified representational space. Preprint at *bioRxiv* https://doi.org/10.1101/2022.09.06.506403 (2022).
184. Geadah, V., Horoi, S., Kerg, G., Wolf, G. & Lajoie, G. Goal-driven optimization of single-neuron properties in artificial networks reveals regularization role of neural diversity and adaptation. Preprint at *bioRxiv* https://doi.org/10.1101/2022.04.29.489963 (2022).
185. Elsayed, G., Ramachandran, P., Shlens, J. & Kornblith, S. Revisiting spatial invariance with low-rank local connectivity. in *International Conference on Machine Learning* 2868–2879 (PMLR, 2020).
186. Zaadnoordijk, L., Besold, T. R. & Cusack, R. Lessons from infant learning for unsupervised machine learning. *Nat. Mach. Intell.* **4**, 510–520 (2022).

187. Rane, S. et al. Predicting word learning in children from the performance of computer vision systems. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2207.09847 (2022).
188. Cadena, S. A. et al. How well do deep neural networks trained on object recognition characterize the mouse visual system? In *Neuro-AI Workshop at the Neural Information Processing Conference* (NeurIPS, 2019).
189. Cao, R. & Yamins, D. Explanatory models in neuroscience: part 2 — constraint-based intelligibility. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2104.01489 (2021).
190. Kanwisher, N., Khosla, M. & Dobs, K. Using artificial neural networks to ask 'why' questions of minds and brains. *Trends Neurosci.* **46**, 240–254 (2023).
191. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
192. Cichy, R. M., Khosla, A., Pantazis, D. & Oliva, A. Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage* **153**, 346–358 (2017).
193. Eickenberg, M., Gramfort, A., Varoquaux, G. & Thirion, B. Seeing it all: convolutional network layers map the function of the human visual system. *NeuroImage* **152**, 184–194 (2017).
194. Averbeck, B. B. Pruning recurrent neural networks replicates adolescent changes in working memory and reinforcement learning. *Proc. Natl Acad. Sci. USA* **119**, e2121331119 (2022).
195. Rust, N. C. & Jannuzi, B. G. Identifying objects and remembering images: insights from deep neural networks. *Curr. Dir. Psychol. Sci.* **31**, 09637214221083663 (2022).
196. Tanaka, H. et al. From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. *Adv. Neural Inf. Process. Syst.* https://papers.nips.cc/paper_files/paper/2019/hash/eeaebbffb5d29ff62799637fc51adb7b-Abstract.html (2019).
197. Berner, J., Grohs, P., Kutyniok, G. & Petersen, P. The modern mathematics of deep learning. in *Mathematical Aspects of Deep Learning* (eds Grohs, P. & Kutyniok, G.) 1–111 (Cambridge Univ. Press, 2022); https://doi.org/10.1017/9781009025096.002.
198. Olshausen, B. A. & Field, D. J. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vis. Res.* **37**, 3311–3325 (1997).
199. Nakkiran, P. et al. Deep double descent: where bigger models and more data hurt. *J. Stat. Mech. Theory Exp.* **2021**, 124003 (2021).
200. Jacot, A., Gabriel, F. & Hongler, C. Neural tangent kernel: convergence and generalization in neural networks. in *Advances in Neural Information Processing Systems* Vol. 31 (NIPS, 2018).
201. Simsek, B. et al. Geometry of the loss landscape in overparameterized neural networks: symmetries and invariances. in *International Conference on Machine Learning* 9722–9732 (PMLR, 2021).
202. Minh, D., Wang, H. X., Li, Y. F. & Nguyen, T. N. Explainable artificial intelligence: a comprehensive review. *Artif. Intell. Rev.* **55**, 3503–3568 (2022).
203. Kar, K., Kornblith, S. & Fedorenko, E. Interpretability of artificial neural network models in artificial intelligence versus neuroscience. *Nat. Mach. Intell.* **4**, 1065–1067 (2022).
204. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1312.6034 (2013).
205. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. in *European Conference on Computer Vision* 818–833 (Springer, 2014).
206. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why should I trust you?' Explaining the predictions of any classifier. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (ACM, 2016).
207. Fong, R. C. & Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. in *Proceedings of the IEEE International Conference on Computer Vision* 3429–3437 (IEEE, 2017).
208. Olah, C., Mordvintsev, A. & Schubert, L. Feature visualization. *Distill* **2**, e7 (2017).
209. Hendricks, L. A. et al. Generating visual explanations. in *European Conference on Computer Vision* 3–19 (Springer, 2016).
210. Herzog, M. H. & Manassi, M. Uncorking the bottleneck of crowding: a fresh look at object recognition. *Curr. Opin. Behav. Sci.* **1**, 86–93 (2015).
211. Doerig, A. et al. Beyond Bouma's window: how to explain global aspects of crowding? *PLOS Comput. Biol.* **15**, e1006580 (2019).
212. Herzog, M. H., Sayim, B., Chicherov, V. & Manassi, M. Crowding, grouping, and object recognition: a matter of appearance. *J. Vis.* **15**, 5–5 (2015).
213. Sabour, S., Frosst, N. & Hinton, G. E. Dynamic routing between capsules. in *Advances in Neural Information Processing Systems* 3856–3866 (NIPS, 2017).
214. Bornet, A., Doerig, A., Herzog, M. H., Francis, G. & Van der Burg, E. Shrinking Bouma's window: how to model crowding in dense displays. *PLoS Comput. Biol.* **17**, e1009187 (2021).
215. Choung, O.-H., Bornet, A., Doerig, A. & Herzog, M. H. Dissecting (un) crowding. *J. Vis.* **21**, 10 (2021).
216. Spoerer, C. J., McClure, P. & Kriegeskorte, N. Recurrent convolutional neural networks: a better model of biological object recognition. *Front. Psychol.* **8**, 1551 (2017).
217. Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behaviour. *Nat. Neurosci.* **22**, 974 (2019).
218. van Bergen, R. S. & Kriegeskorte, N. Going in circles is the way forward: the role of recurrence in visual inference. *Curr. Opin. Neurobiol.* **65**, 176–193 (2020).
219. Kreiman, G. & Serre, T. Beyond the feedforward sweep: feedback computations in the visual cortex. *Primates* **9**, 16 (2019).
220. Nayebi, A. et al. Recurrent connections in the primate ventral visual stream mediate a trade-off between task performance and network size during core object recognition. *Neural Comput.* **34**, 1652–1675 (2022).
221. Sullivan, J., Mei, M., Perfors, A., Wojcik, E. & Frank, M. C. SAYCam: a large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open. Mind* **5**, 20–29 (2021).
222. Clay, V., König, P., Kühnberger, K.-U. & Pipa, G. Learning sparse and meaningful representations through embodiment. *Neural Netw.* **134**, 23–41 (2021).
223. Gan, C. et al. The threeDworld transport challenge: a visually guided task-and-motion planning benchmark for physically realistic embodied AI. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2103.14025 (2021).
224. Chen, Y. et al. COCO-Search18 fixation dataset for predicting goal-directed attention control. *Sci. Rep.* **11**, 8776 (2021).
225. Zhuang, C. et al. Unsupervised neural network models of the ventral visual stream. *Proc. Natl Acad. Sci. USA* **118**, e2014196118 (2021).
226. Konkle, T. & Alvarez, G. A. A self-supervised domain-general learning framework for human ventral stream representation. *Nat. Commun.* **13**, 491 (2022).
227. Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C. & Richards, B. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. in *Advances in Neural Information Processing Systems* Vol. 34 (NIPS, 2021).
228. Nayebi, A. et al. Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. Preprint at *bioRxiv* https://doi.org/10.1101/2021.06.16.448730 (2022).
229. Mineault, P., Bakhtiari, S., Richards, B. & Pack, C. Your head is there to move you around: goal-driven models of the primate dorsal pathway. in *Advances in Neural Information Processing Systems* Vol. 34 (NIPS, 2021).
230. Stringer, S. M., Rolls, E. T. & Trappenberg, T. P. Self-organizing continuous attractor network models of hippocampal spatial view cells. *Neurobiol. Learn. Mem.* **83**, 79–92 (2005).
231. Tsodyks, M. Attractor neural network models of spatial maps in hippocampus. *Hippocampus* **9**, 481–489 (1999).
232. Uria, B. et al. The spatial memory pipeline: a model of egocentric to allocentric understanding in mammalian brains. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.11.378141 (2020).
233. Whittington, J. C. et al. The Tolman–Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell* **183**, 1249–1263.e23 (2020).
234. Whittington, J. C., Warren, J. & Behrens, T. E. Relating transformers to models and neural representations of the hippocampal formation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2112.04035 (2021).
235. Acunzo, D. J., Low, D. M. & Fairhall, S. L. Deep neural networks reveal topic-level representations of sentences in medial prefrontal cortex, lateral anterior temporal lobe, precuneus, and angular gyrus. *NeuroImage* **251**, 119005 (2022).
236. Riveland, R. & Pouget, A. A neural model of task compositionality with natural language instructions. Preprint at *bioRxiv* https://doi.org/10.1101/2022.02.22.481293 (2022).
237. Xu, P., Zhu, X. & Clifton, D. A. Multimodal learning with transformers: a survey. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2206.06488 (2022).
238. Ivanova, A. A. et al. Beyond linear regression: mapping models in cognitive neuroscience should align with research goals. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2208.10668 (2022).
239. Peterson, J. C., Abbott, J. T. & Griffiths, T. L. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cogn. Sci.* **42**, 2648–2669 (2018).
240. Golan, T., Raju, P. C. & Kriegeskorte, N. Controversial stimuli: pitting neural networks against each other as models of human cognition. *Proc. Natl Acad. Sci. USA* **117**, 29330–29337 (2020).
241. Geirhos, R., Meding, K. & Wichmann, F. A. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Adv. Neural Inf. Process. Syst.* **33**, 13890–13902 (2020).
242. Biscione, V. & Bowers, J. S. Do DNNs trained on natural images acquire Gestalt properties? Preprint at *arXiv* https://doi.org/10.48550/arXiv.2203.07302 (2022).
243. Feather, J., Durango, A., Gonzalez, R. & McDermott, J. Metamers of neural networks reveal divergence from human perceptual systems. *Advances in Neural Information Processing Systems* Vol. 32 (NIPS, 2019).
244. Mastrogiuseppe, F. & Ostojic, S. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron* **99**, 609–623.e29 (2018).
245. Dujmović, M., Bowers, J., Adolfi, F. & Malhotra, G. The pitfalls of measuring representational similarity using representational similarity analysis. Preprint at *bioRxiv* https://doi.org/10.1101/2022.04.05.487135 (2022).
246. Elmoznino, E. & Bonner, M. F. High-performing neural network models of visual cortex benefit from high latent dimensionality. Preprint at *bioRxiv* https://doi.org/10.1101/2022.07.13.499969 (2022).
247. Schaeffer, R., Khona, M. & Fiete, I. R. No free lunch from deep learning in neuroscience: a case study through models of the entorhinal-hippocampal circuit. in *ICML 2022 2nd AI for Science Workshop* (ICML, 2022).
248. Crick, F. The recent excitement about neural networks. *Nature* **337**, 129–132 (1989).
249. Szegedy, C. et al. Intriguing properties of neural networks. in *2nd International Conference on Learning Representations, ICLR 2014* (ICLR, 2014).

250. Moosavi-Dezfooli, S.-M., Fawzi, A. & Frossard, P. Deepfool: a simple and accurate method to fool deep neural networks. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2574–2582 (IEEE, 2016).
251. Nguyen, A., Yosinski, J. & Clune, J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 427–436 (IEEE, 2015).
252. Baker, N., Lu, H., Erlikhman, G. & Kellman, P. J. Deep convolutional networks do not classify based on global object shape. *PLoS Comput. Biol.* **14**, e1006613 (2018).
253. Heinke, D., Wachman, P., van Zoest, W. & Leek, E. C. A failure to learn object shape geometry: implications for convolutional neural networks as plausible models of biological vision. *Vis. Res.* **189**, 81–92 (2021).
254. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1412.6572 (2014).
255. Bai, T., Luo, J., Zhao, J., Wen, B. & Wang, Q. Recent advances in adversarial training for adversarial robustness. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2102.01356 (2021).
256. Dapello, J. et al. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. *Adv. Neural Inf. Process. Syst.* **33**, 13073–13087 (2020).
257. Malhotra, G., Evans, B. D. & Bowers, J. S. Hiding a plane with a pixel: examining shape-bias in CNNs and the benefit of building in biological constraints. *Vis. Res.* **174**, 57–68 (2020).
258. Machiraju, H., Choung, O.-H., Herzog, M. H. & Frossard, P. Empirical advocacy of bio-inspired models for robust image recognition. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2205.09037 (2022).
259. Ilyas, A. et al. Adversarial examples are not bugs, they are features. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1905.02175 (2019).
260. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
261. Elsayed, G. et al. Adversarial examples that fool both computer vision and time-limited humans. in *Advances in Neural Information Processing Systems* 3910–3920 (NIPS, 2018).
262. Guo, C. et al. Adversarially trained neural representations are already as robust as biological neural representations. in *International Conference on Machine Learning* 8072–8081 (PMLR, 2022).
263. Zhou, Z. & Firestone, C. Humans can decipher adversarial images. *Nat. Commun.* **10**, 1334 (2019).
264. Hermann, K., Chen, T. & Kornblith, S. The origins and prevalence of texture bias in convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **33**, 19000–19015 (2020).
265. Evans, B. D., Malhotra, G. & Bowers, J. S. Biological convolutions improve DNN robustness to noise and generalisation. *Neural Netw.* **148**, 96–110 (2022).
266. Geirhos, R. et al. Partial success in closing the gap between human and machine vision. in *Advances in Neural Information Processing Systems* Vol. 34 (NIPS, 2021).
267. Jagadeesh, A. V. & Gardner, J. L. Texture-like representation of objects in human visual cortex. *Proc. Natl Acad. Sci. USA* **119**, e2115302119 (2022).
268. Fodor, J. A. & Pylyshyn, Z. W. Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**, 3–71 (1988).
269. Jackendoff, R. Précis of foundations of language: brain, meaning, grammar, evolution. *Behav. Brain Sci.* **26**, 651–665 (2003).
270. Marcus, G. F. *The Algebraic Mind: Integrating Connectionism and Cognitive Science* (MIT Press, 2003).
271. Quilty-Dunn, J., Porot, N. & Mandelbaum, E. The best game in town: the re-emergence of the language of thought hypothesis across the cognitive sciences. *Behav. Brain Sci.* https://doi.org/10.1017/S0140525X22002849 (2022).
272. Chomsky, N. *Language and Mind* (Cambridge Univ. Press, 2006).
273. Frankland, S. M. & Greene, J. D. Concepts and compositionality: in search of the brain's language of thought. *Annu. Rev. Psychol.* **71**, 273–303 (2020).
274. Pinker, S. & Prince, A. On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition* **28**, 73–193 (1988).
275. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **2**, 359–366 (1989).
276. Santoro, A., Lampinen, A., Mathewson, K., Lillicrap, T. & Raposo, D. Symbolic behaviour in artificial intelligence. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2102.03406 (2021).
277. Mul, M., Bouchacourt, D. & Bruni, E. Mastering emergent language: learning to guide in simulated navigation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1908.05135 (2019).
278. ChatGPT: optimizing language models for dialogue. OpenAI https://openai.com/blog/chatgpt/ (2022).
279. Shahriar, S. & Hayawi, K. Let's have a chat! A conversation with ChatGPT: technology, applications, and limitations. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2302.13817 (2023).
280. OpenAI. GPT-4 technical report. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.08774 (2023).
281. Hinton, G. How to represent part-whole hierarchies in a neural network. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2102.12627 (2021).
282. Higgins, I. et al. beta-vae: learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations* https://openreview.net/forum?id=Sy2fzU9gl (2017).
283. Higgins, I. et al. Towards a definition of disentangled representations. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1812.02230 (2018).
284. Eslami, S. A. et al. Neural scene representation and rendering. *Science* **360**, 1204–1210 (2018).
285. Graves, A., Wayne, G. & Danihelka, I. Neural turing machines. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1410.5401 (2014).
286. Garnelo, M., Arulkumaran, K. & Shanahan, M. Towards deep symbolic reinforcement learning. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1609.05518 (2016).
287. Holyoak, K. J. The proper treatment of symbols. in *Cognitive Dynamics: Conceptual and Representational Change in Humans and Machines* Vol. 229 (Psychology Press, 2000).
288. Smolensky, P., McCoy, R. T., Fernandez, R., Goldrick, M. & Gao, J. Neurocompositional computing: from the central paradox of cognition to a new generation of AI systems. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2205.01128 (2022).
289. Hummel, J. E. Getting symbols out of a neural architecture. *Connect. Sci.* **23**, 109–118 (2011).
290. Smolensky, P. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artif. Intell.* **46**, 159–216 (1990).
291. Eliasmith, C. *How to Build a Brain: A Neural Architecture for Biological Cognition* (Oxford Univ. Press, 2013).
292. Flesch, T., Juechems, K., Dumbalska, T., Saxe, A. & Summerfield, C. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron* **110**, 1258–1270 (2022).
293. Molano-Mazon, M. et al. NeuroGym: an open resource for developing and sharing neuroscience tasks. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/aqc9n (2022).
294. Koulakov, A., Shuvaev, S., Lachi, D. & Zador, A. Encoding innate ability through a genomic bottleneck. Preprint at *bioRxiv* https://doi.org/10.1101/2021.03.16.435261 (2022).
295. Heinke, D. Computational modelling in behavioural neuroscience: methodologies and approaches (minutes of discussions at the workshop in Birmingham, UK, in May 2007). in *Computational Modelling in Behavioural Neuroscience* 346–352 (Psychology Press, 2009).
296. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154 (1962).
297. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025 (1999).
298. Wen, H. et al. Neural encoding and decoding with deep learning for dynamic natural vision. *Cereb. Cortex* **28**, 4136–4160 (2018).
299. Popper, K. *The Logic of Scientific Discovery* (Routledge, 2005).
300. Duhem, P. M. M. *The Aim and Structure of Physical Theory* Vol. 13 (Princeton Univ. Press, 1991).
301. Duhem, P. Physical theory and experiment. in *Can Theories Be Refuted?* 1–40 (Springer, 1976).
302. Gillies, D. Philosophy of science in the twentieth century: four central themes. *Br. J. Philos. Sci.* **45**, 1066–1069 (1994).
303. Quine, W. v. O. Two dogmas of empiricism. in *Can theories Be refuted?* 41–64 (Springer, 1976).
304. Kuhn, T. S. *The Structure of Scientific Revolutions* (Univ. Chicago Press, 2012).

Acknowledgements
The authors acknowledge support by the SNF grant 203018 (A.D.), the ERC stg grant 101039524 TIME (T.C.K.) and the Max Planck Research Group grant of Martin N. Hebart (K.S.).

Author contributions
A.D., R.P.S., K.S. and T.C.K. initiated the project and wrote the first draft of the article. A.D., R.P.S., K.S., B.R., J.I., G.W.L., T.K., M.A.J.v.G. and T.C.K. contributed significantly to subsequent versions of this manuscript. All authors researched data for the article and contributed substantially to the conceptualization of the research programme.

Competing interests
The authors declare no competing interests.

Additional information
Peer review information *Nature Reviews Neuroscience* thanks Gemma Roig, who co-reviewed with Martina Vilas; Benjamin Cowley; Dietmar Heinke; and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.