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.

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

Nature Reviews Neuroscience | Volume 24 | July 2023 | 431–450
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. 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.

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’. This approach has merit. Its controlled settings allow for maximal interpretability and suggest a clear taxonomy of neural selectivity. 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. 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 neurocomputational 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 (ANNs) make them a model class well suited to tackle these challenges, with deep neural networks 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 fulfill ‘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 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. Furthermore, the more recent focus on multilevel understanding of brain function of this approach goes beyond classic connectionist models of the twentieth century, 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 neuroscience settings, including vision, audition, semantics, language, reading, decision-making, attention, memory, game playing, motor control and the formation and coding principles of brain areas (reviewed elsewhere), further demonstrating that the larger neuroconnectionism research community involves, among many other areas, sensory
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. 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, 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, 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.

**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).

**Neuroconnectionism as a Lakatosian research programme**

In the philosophy of science, Lakatos 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 fidelity 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).
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 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, and later reinforced by Quine (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

Where (A1, and A2, and A3, ..., and An) 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. Quine 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...
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, 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
approaches that share distinctive features. Hence, the member projects of neuroconnectionism are best understood as constituting a diverse but cohesive family of resemblance\textsuperscript{73}. Hence, the member projects of neuroconnectionism are best understood as constituting a diverse but cohesive family of approaches that share distinctive features.

Perspective

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 habitable 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.

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\textsuperscript{4,29}, 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\textsuperscript{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\textsuperscript{25}. 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\textsuperscript{29}. 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\textsuperscript{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\textsuperscript{44} and from approaches using biophysical models\textsuperscript{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. 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 single 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 incontemporaneous 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. 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. 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 instantiated, 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. 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, convolutional layers 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). 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, richer rate-based neurons, spiking neurons and neurons with multiple compartments, 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, leveraging recent efforts to record large neural data sets leveraging recent efforts to record large neural data sets, 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 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, predictive coding, image generation, temporal stability and energy efficiency) and behavioural reward. Different objectives impact what the network learns and thereby the ability of the network to model different brain areas. 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. 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. However, backpropagation in its standard form is not biologically plausible, but multiple avenues of research seek to define backpropagation in biologically more plausible ways. Other learning rules
various complementary approaches are needed. Thus, developing length Importantly, no single-model testing method is perfect and constraints the brain faces such as metabolic costs and wiring from neural data to behavioural, ideally while considering physical

Although no single-model testing method is perfect, constraints the brain faces such as metabolic costs and wiring from neural data to behavioural outcomes, ideally while considering physical and computational limitations, the most common methods are representational similarity analysis (comparing representational geometries of population activities) and encoding models (predicting brain activity from ANNs units via linear regression). 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. Hence, to ensure that an ANN

As with the other approaches used to cast coarse-grained neuroscientific hypotheses into mathematically precise neuroconnectionist 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 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). 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 and wiring length. 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, board games or video games. 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, detailed analysis of error patterns, testing on out-of-distribution examples and reproducing psychophysical results that target particular aspects of processing (reviewed elsewhere 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). 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

Perspective

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 ANNs 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).
the agreement between the ANN and the brain. A direct comparison of representational geometries between brain data and/or ANNs 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. 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 and clarifying which aspects of brain computation should be targeted using RSA.

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. Here, the activity of each biological unit (neuron or voxel) is predicted as a linear combination of ANN unit activations. This, hence, 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. 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.

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. 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. 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. 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 classical experimentation. These in silico experiments include reliance on network initialization and tests for the emergence of brain-like computations in individual network units, selectivity profiles, and cell types. 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. 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 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 (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, 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.

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. 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 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). 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. 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...
Successful predictions of neuronal firing rates from ANNs (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 does18–21. This is done by optimizing a model to fulfill one (or many) pre-defined objective(s)19, 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 coding119, 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 responses195–197. 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 fulfill198. Thus, this normative modelling result suggests that face patches carry out computations with disentanglement as a goal. Recently, optimizing models to fulfill 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 networks199. 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 space200. 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 development201. Further examples proposed to explain adolescent changes in working memory by pruning in ANNs202 and predicted the memorability of images in humans from response magnitude in ANN layers203.

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 voxels279–281. 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 phenomena204. 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 images205. 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’206. Insights from deep mathematics are of great importance for understanding complex neural processes, as the brain, too, is highly overparameterized207. 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 insight208. 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 layer209. 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 conditions210. 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 minimum211. 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 elsewhere202). Many of these deep
Perspective

**a**
Neocognitron
Visual system mechanism proposed by Fukushima, based on the findings by Hubel and Wiesel

**b**
HMAX and IT
The neocognitron-based HMAX model mimics primate ventral stream receptive field sizes and accounts for human rapid animacy categorization behaviour

**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

**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

**g**
Neural dynamics
Feedforward CNNs cannot account for neural 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
learning approaches can also be used in brain science\textsuperscript{203}. Feature visualization methods investigate what ANNs detect by quantifying which parts of the input most strongly impact activities and network behaviour\textsuperscript{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)\textsuperscript{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\textsuperscript{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 psychophysiological 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\textsuperscript{210–212}, 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\textsuperscript{213,214} 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\textsuperscript{215}, 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\textsuperscript{216–218}. 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\textsuperscript{20,21,23,24}. 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\textsuperscript{25}. Adding recurrence to networks improved matching both behavioural data and neural activity patterns that occurred with longer delays\textsuperscript{21,23,99,100}. In addition, researchers explored ways to improve the training data sets beyond computer vision benchmarks, including a more ecologically relevant selection of object categories\textsuperscript{19}, video data\textsuperscript{20}, embodiment\textsuperscript{20,23,24} and goal-directed eye movements\textsuperscript{20,24}. 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\textsuperscript{20,21,23,24,25}. Moreover, self-supervised training on video data can account for the functional distinction between the dorsal and ventral pathways in the brain\textsuperscript{21}, and self-supervised training on lower-resolution inputs provides a better fit to electrophysiological data of mouse visual cortex\textsuperscript{23}. 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\textsuperscript{20} and by self-motion-related loss functions\textsuperscript{23,24}.

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 theiet of the research programme evolved. Early ANN-based hippocampal models were often attractor networks\textsuperscript{23,24,25}, 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\textsuperscript{21,23,24}. 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\textsuperscript{23,24}, and other brain circuits, such as the mnemonic circuits of the medial temporal lobes\textsuperscript{23,24}. 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\textsuperscript{23,24,25}.

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\textsuperscript{20,23,24,25} and training networks end-to-end directly to match neural data\textsuperscript{19,20,100–111}. Further examples include work highlighting the important individual variability across network instances\textsuperscript{20,112}, explicitly pitting networks against each other using ‘controversial stimuli’\textsuperscript{210} and integrating hierarchical\textsuperscript{19} and temporal\textsuperscript{113} aspects of information processing into model comparisons. In addition, performing detailed psychophysical experiments\textsuperscript{211–213} using metamers to compare humans and models\textsuperscript{214}, and developing interpretable low-rank recurrent networks\textsuperscript{115} 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\textsuperscript{108–111,215}, this progress 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\textsuperscript{216,217}. Although capable of impressive behaviours, current ANNs do not model all aspects of biological behaviour equally well: they sometimes generalize poorly\textsuperscript{218}, can be easily fooled\textsuperscript{219–221} and behave differently from humans in many psychophysical settings\textsuperscript{222–224}. Perhaps, the most famous of these differences is given by adversarial examples\textsuperscript{225–227}, 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\textsuperscript{228}. 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\textsuperscript{229,230} be leveraged to increase robustness? Biologically inspired features have had some success increasing robustness, but work remains to fully understand their impact (reviewed elsewhere\textsuperscript{231}). Third, where do adversarial examples come from? Perhaps, the most influential view is that they are features, not bugs\textsuperscript{232}: 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\textsuperscript{233} 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
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. 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 for example in which texture randomization had negligible effects on the global processing abilities of CNNs, improving data augmentation, using a more biologically inspired first network layer or training ANNs on substantially larger data sets. This shape versus texture distinction was studied in the brain too, revealing that inferior temporal cortex might rely on a textural encoding, 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. In addition, owing to the human capacity to understand and create novel combinations of concepts, ANNs are thought to run into tractability or scalability issues. Yet, even ardent defenders of symbolism agree that ANNs should, in principle, be able to implement the desired symbolic systems, as such a capacity is an immediate result of the universal approximation theorems that hold for multilayer and recurrent neural networks. 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 neuropsychological 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. 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, which have helped bring models ever closer to human reasoning (with limitations such as fact hallucination and reasoning errors). 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. 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).

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 foiling 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 (ii), a focus on embodiment rather than treating networks as artificial brains in vats, (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) 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, 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., Maclver, M. A. & Poegel, 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. Servos, J., Ohta, S. & MacDonald, B. Functional neuroanatomy of face and object processing: a position emission tomography study. Brain 115, 35–36 (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. Tso, D. Y., Freiwald, W. A., Knutson, 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. Rev. Neurosci. 8, 1647–1650 (2005).
8. Vinken, K., Konkle, T. & Livingstone, M. The neural code for ‘face cells’ is not face specific. Nat. Rev. Neurosci. 17, 116–126 (2016).
9. Watanabe, T., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature 503, 78–84 (2013).
10. Quax, S. & van Gerven, M. Emergent mechanisms of evidence integration in recurrent neural networks. PLoS ONE 13, e0205676 (2018).
11. Doerr, A. R. & Brookes, G. K. Why biological attention mechanisms improve task performance in a large-scale visual system model. eLife 7, e38105 (2018).
12.ð Hampson, T. et al. Diverse task-driven modeling of macaque V4 reveals functional specialization across semantic tasks. Preprint at bioRxiv https://doi.org/10.1101/2022.05.18.492503 (2022).
13. 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).
14. Saxe, A. M., McClelland, J. L. & Gangu, S. A mathematical theory of semantic development in deep neural networks. Proc. Natl Acad. Sci. USA 116, 1537–1546 (2019).
15. Doerr, A. et al. Semantic scene descriptions as an objective of human vision. Preprint at arXiv https://doi.org/10.48550/arXiv.2209.11737 (2022).
16. Caucheteux, C. & King, J. R. Brains and algorithms partially converge in natural language processing. Commun. Biol. 5, 134 (2022).
17. Schirmpf, 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 (2022).
18. 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, e201479118 (2021).
19. Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J. & Koth-Kumar, Z. Deep reinforcement learning and its neuroscientific implications. Neuron 107, 603–616 (2020).
20. Dabney, W. et al. A functional code for value in dopamine-based reinforcement learning. Nature 577, 671–675 (2020).
21. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature 503, 78–84 (2013).
22. Quax, S. & van Gerven, M. Emergent mechanisms of evidence integration in recurrent neural networks. PLoS ONE 13, e0205676 (2018).
23. Henderson, P. J., Seth, A. K., Petrov, O. & Marois, R. The emergence of topographic organization in a non-convolutional deep neural network (Neuromatch 40). Perception 51, 74–75 (2022).
24. Allen, K. M. et al. Unsupervised deep learning identifies semantic disentanglement in single-layered frontoparietal face processing neurons. Nat. Commun. 12, 4456 (2021).
25. 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/10.1101/2020.07.09.185166 (2020).
