If deep learning is the answer, what is the question?

Andrew Saxe, Stephanie Nelli and Christopher Summerfield

Abstract | Neuroscience research is undergoing a minor revolution. Recent advances in machine learning and artificial intelligence research have opened up new ways of thinking about neural computation. Many researchers are excited by the possibility that deep neural networks may offer theories of perception, cognition and action for biological brains. This approach has the potential to radically reshape our approach to understanding neural systems, because the computations performed by deep networks are learned from experience, and not endowed by the researcher. If so, how can neuroscientists use deep networks to model and understand biological brains? What is the outlook for neuroscientists who seek to characterize computations or neural codes, or who wish to understand perception, attention, memory and executive functions? In this Perspective, our goal is to offer a road map for systems neuroscience research in the age of deep learning. We discuss the conceptual and methodological challenges of comparing behaviour, learning dynamics and neural representations in artificial and biological systems, and we highlight new research questions that have emerged for neuroscience as a direct consequence of recent advances in machine learning.

Recent years have seen a dramatic resurgence in optimism about the progress of artificial intelligence (AI) research, driven by advances in deep learning[1]. 'Deep learning' is the name given to a methodological toolkit for building multilayer (or 'deep') neural networks that solve challenging problems in supervised classification[2], generative modelling[3] or reinforcement learning[4,5]. Neuroscience and AI research have a rich shared history[6], and deep networks are now increasingly being considered as promising theories of neural computation. The recent literature is studded with comparisons of the behaviour and activity of biological and artificial systems[7–21], summarized in a growing number of review articles[22–30].

In this Perspective, we assess the opportunities and challenges presented by this new wave of intellectual synergy between neuroscience and AI research. We begin by considering the recent proposals that have sought to reframe neural theory as a deep learning problem. We assess extant evidence that deep networks form representations or exhibit behaviours in ways that resemble biological agents and consider a host of new questions, inspired by deep learning, that neuroscientists are only just beginning to address. In doing so, we highlight specific falsifiable hypotheses that often underpin deep learning models, drawing on the domains of perception, memory, inference and control processes. We point to the limits of correlating representations of brains and complex deep learning architectures, and argue for a focus on learning trajectories and complex behaviours. Finally, we ask how deep network theories can provide explanation and understanding, by drawing on recent research that is beginning to develop mathematical descriptions of network learning dynamics and behaviour. In doing so, we argue that deep networks can and should be used to provide a new generation of falsifiable theories of how humans and other animals think, learn and behave.

Neoconnectionism?

The idea that neural networks can serve as theories of neural computation is not new. During the parallel distributed processing movement of the 1980s, psychologists and computer scientists proposed neural networks as solutions to key problems in perception, memory and language[31]. Contemporary deep networks resemble scaled-up connectionist models, and recent advances in machine learning are also heavily indebted to the ubiquity of digital data and the relatively low cost of computation in the twenty-first century[26]. It might thus be tempting to dismiss current excitement around deep learning models for neuroscience as a rehashing of earlier ideas, owing more to the slow churn of scientific fashion than to genuine intellectual progress. However, many researchers believe that deep learning models have the potential to radically reshape neural theory, and to open new avenues for symbiotic research between neuroscience and AI research[23,32–34]. This is because contemporary deep networks are grounded in quasi-naturalistic sensory signals (such as image pixels[13] or auditory spectrograms[15]) that allow them to perform tasks of far greater complexity than was previously possible. Contemporary deep networks can thus learn 'end-to-end' (that is, without researcher intervention) in a sensory ecology that resembles our own: natural sounds and scenes for supervised learning and generative modelling, and 3D environments with realistic physics for deep reinforcement learning. This advent of end-to-end models of biological function has enabled researchers to attempt to model, for the first time, the de novo emergence of neural computations that can solve real-world problems.

Networks capable of high performance on complex real-world tasks have enabled a host of recent advances at the intersection of machine learning and neuroscience. For example, one major line of research has examined the representations formed by supervised deep networks that are trained to label objects in natural scenes[2] (FIG. 1). A striking observation is that biologically plausible neural representations can emerge in networks that combine gradient descent with a handful of simple computational principles[29] (gradient descent is a training method where weights are adjusted incrementally to encourage the network outputs towards an objective). When deep networks are endowed with properties including local connectivity, convolutions,

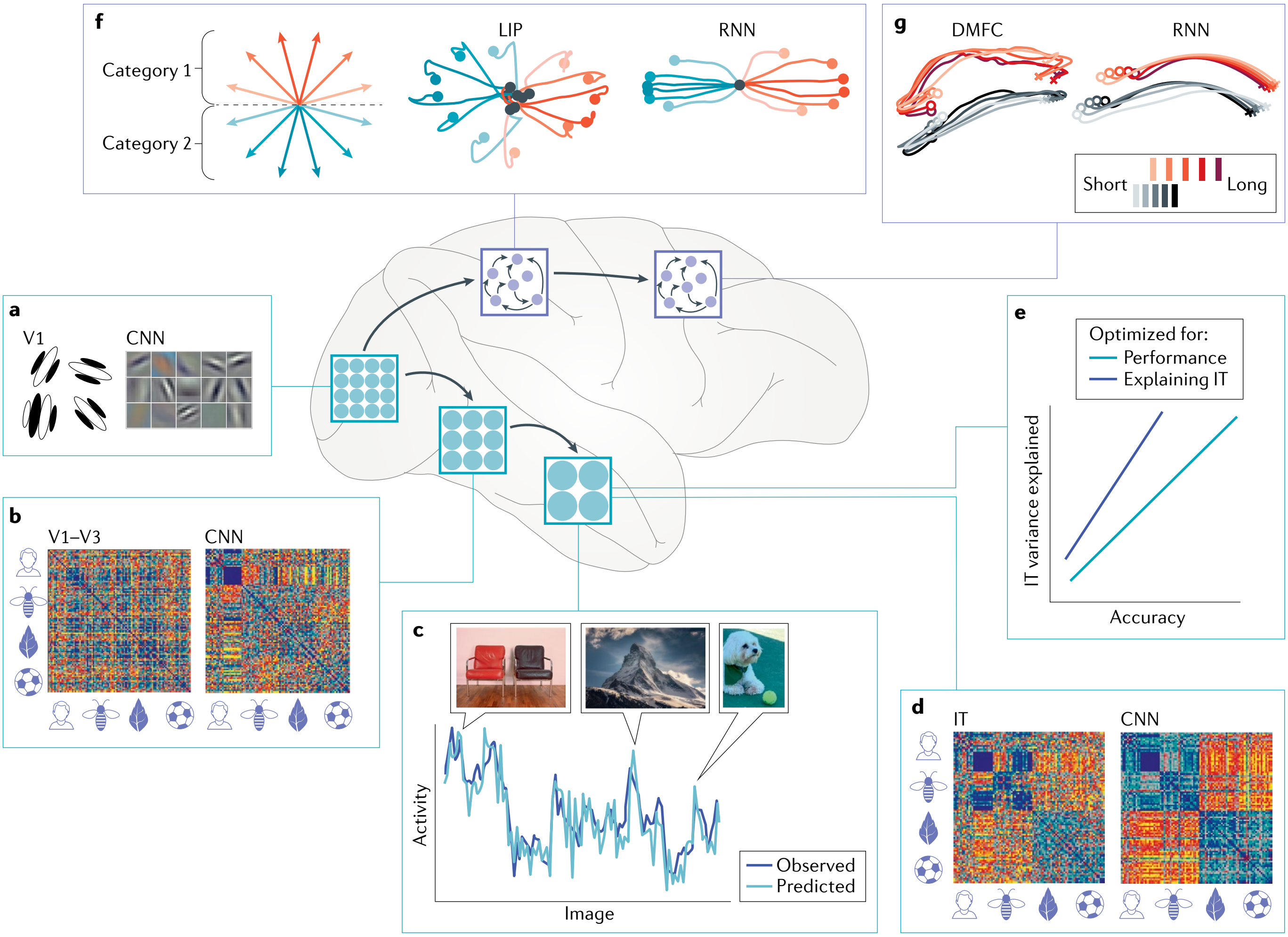


Fig. 1 | **Representational equivalence between neural networks and the primate brain.** This figure summarizes evidence for representational correspondence between deep networks and biological brains. **a** | Left: schematic illustration of simple and complex cell receptive fields from mammalian primary visual cortex (V1). Right: example filters learned in the first hidden layer of a deep convolutional neural network (CNN)[2]. **b** | Representational similarity analysis is a method by which the similarity of the population response to each stimulus (images of faces, bees, leaves and balls in this example) can be assessed. Example representational similarity matrices illustrating the similarity (blue indicating similar and red indicating dissimilar) in population activity evoked by objects in early visual areas of the primate brain (left; recorded with electrophysiology) and in the intermediate layers in a deep CNN[14] (right). **c** | Hypothetical neural firing rates in response to a series of natural images (dark blue trace) and corresponding hypothetical activity predicted as a linear transform of the neural network activity (light blue trace)[13]. **d** | Representational similarity matrices as in part **b** but comparing inferior temporal cortex (IT) with the final layers of a CNN[14]. **e** | Illustration of the relationship between variance explained in IT signals and classification accuracy for pseudo-randomly generated neural networks that are trained to maximize either classification performance (light blue line) or explanation of variance in neural signals (dark blue line)[13]. **f** | Left: state space analysis of neural signals from macaque lateral intraparietal area (LIP) recorded during a dot motion categorization task. Red and blue lines show different motion directions belonging to opposing categories; trajectories are plotted in two dimensions. Right: same analysis conducted on hidden units of a recurrent neural network (RNN)[45]. **g** | Left: state space analysis performed on neural signals recorded from macaque dorsomedial prefrontal cortex (DMPFC) during performance of a long-interval or short-interval reproduction task, plotted in three dimensions. Right: same analysis conducted on hidden units of an RNN[48]. Part **a** adapted with permission from REF.[2], Neural Information Processing Systems. Parts **b** and **d** adapted from REF.[14], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Panel **c** left insert credit: Wolfgang Flamisch/Getty. Panel **c** middle insert credit: KDP/Getty. Part **f** adapted with permission from REF.[45], Elsevier. Part **g** adapted with permission from REF.[48], Elsevier.

pooling and normalization, the early layers acquire simple filters for orientation and spatial frequency[2], just like neurons in the primary visual cortex (FIG. 1a), whereas in deeper layers, the distributions and similarity structure of neural representations for objects and categories resemble those in the primate ventral stream[12–14,19] (FIG. 1b,d). Notably, representational equivalence may be stronger in networks that perform object recognition more accurately[13] (FIG. 1e).

One corollary of these findings is that the sophisticated behaviours and structured neural representations observed in humans and other animals might emerge from a limited set of computational principles, as long as the input data are sufficiently rich and the network is appropriately optimized[25,35].

A deep learning framework

This claim has potentially profound implications for neuroscience. It has already prompted calls for systems neuroscientists to refrain from building theories that impose intuitive functional importance on neural circuits by fiat, and instead to study the computations that emerge spontaneously during the training of deep networks[23,32–34]. One such enjoinder[33] that invokes the term 'deep learning framework' encourages researchers to avoid explicit characterizations of neural computation (for example, how simulated neurons with handcrafted tuning curves and hand-engineered network connectivity might implement a certain function such as object recognition). Instead, it proposes that the role of the researcher is to specify the overall network architecture, the learning rule and the cost function; control is thus relinquished over the microstructure of computation, which instead emerges organically over the course of network training[33]. This proposal has raised the question of whether neural computation is sufficiently interpretable to be worth explaining at all. A related proposal draws an analogy between optimization over computation in neural networks and optimization over biological forms by evolution: in both cases, interpretable functional adaptations emerge without meaningful constraints being imposed on the search process[32]. In other words, it has been claimed that neural systems are fundamentally uninterpretable, and that structured theories of perception and cognition are 'just-so stories' that reflect more closely the researcher's quest for meaning than the reality of neural computation[32]. We consider this view in more detail herein.

The claim has also been made that modelling brains as neural networks relieves researchers of the burden of exhaustively documenting and interpreting the coding properties of single neurons[33]. As methodological advances have permitted simultaneous recordings from large numbers of neurons[36], a doctrine has emerged according to which neural representation is dynamically multiplexed across populations[37]. From this perspective, single neurons code for multiple experimental variables and their interactions[38–41], exhibiting non-linear mixed selectivity. Although a focus on population coding emerged independently of the growing interest in deep learning, mixed selectivity is often (but not always)[42] a hallmark of coding in deep network models[4]. In the brain, this tendency seems to be most pronounced in higher cortical areas, such as the parietal cortex and prefrontal cortex, that support working memory and action selection[38–40]. In these regions, the coding properties of single neurons can be highly heterogenous and vary in mystifying ways over the course of a given trial[38,39,43]. However, when neural activity is examined at the population level — for example, using dimensionality reduction — neural patterns emerge that meaningfully distinguish experimental variables[44,45].

Another key observation is that these patterns of population activity can be recreated when the same analysis is applied to unit activations in recurrent neural networks trained to evaluate time-varying decision evidence[44–46] (FIG. 1f), judge the length of a time interval[47,48] (FIG. 1g) or maintain information over a delay period[49–51]. Accordingly, deep recurrent neural networks that have been trained from scratch are increasingly being proposed as computational theories for sensorimotor integration and working memory that go beyond existing models with more 'handcrafted' features, such as attractor models. In the domain of working memory, a particularly interesting new line of research has used recurrent networks to address a key question in systems neuroscience, namely when codes for stored information should be static or dynamic[49]. This work has contributed to claims that it is futile to characterize the coding properties of individual cells or to infer how they participate in computation[38]. Instead, it is argued the computational model is explainable only at the aggregate level of the population, which is ultimately driven by the structure of the network and the way it is optimized.

Together, these findings have been invoked to argue that attempts to explain computation in single neurons or local brain areas are fruitless, and that meaningful descriptions of neural computation are better given by design choices or hyperparameter settings from machine learning models. For our part, although we celebrate the opportunity to study biological brains through the lens of neural network models, we reject the view that this means giving up on the attempt to explain computation. We elaborate on this in the following sections.

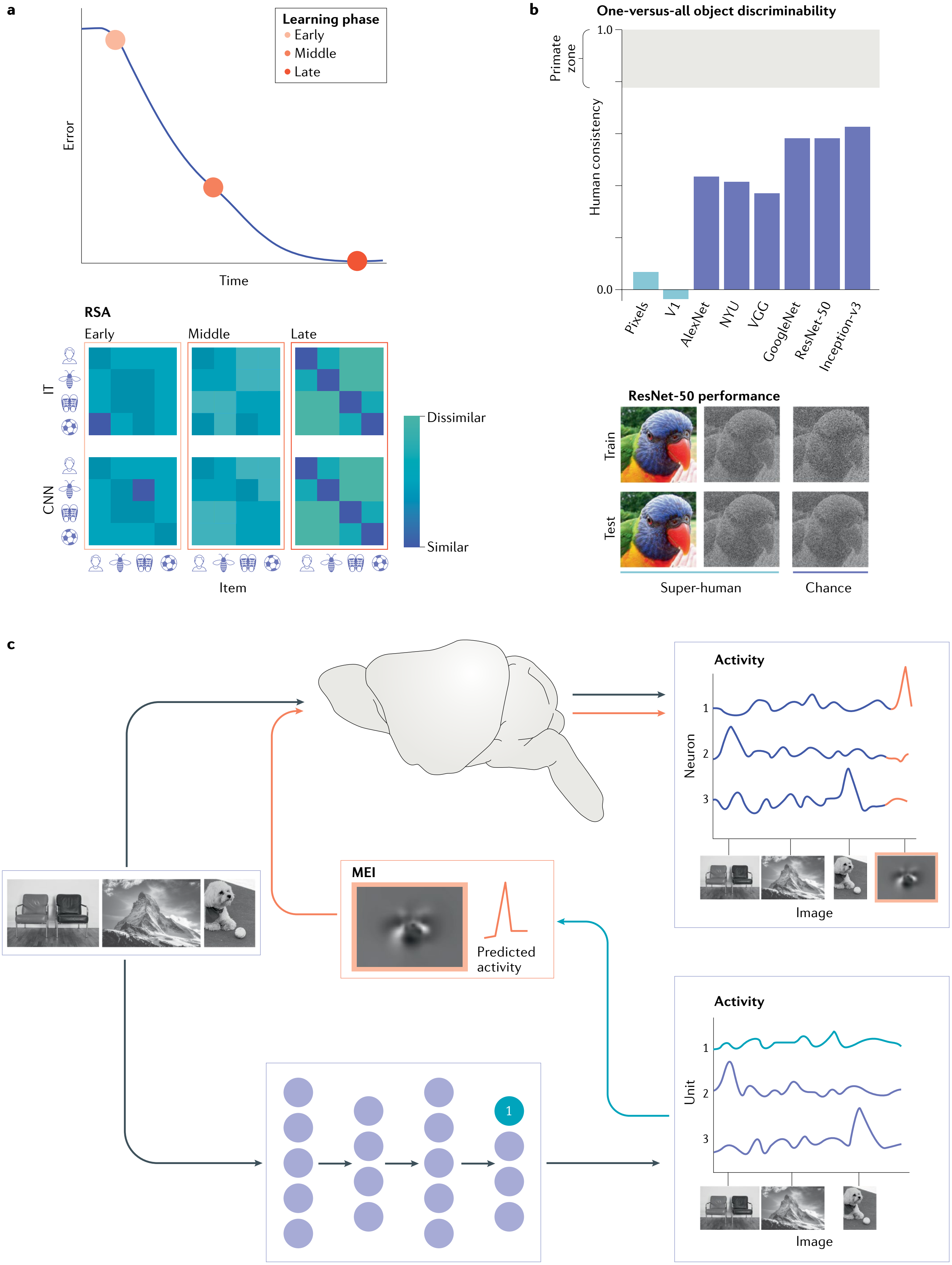
From framework to hypothesis

The deep learning framework proposes powerful new tools for modelling the bewildering volumes of data that are now routinely recorded in systems neuroscience laboratories. However, we hope that enthusiasm for deep networks as computational models will be tempered with a sober consideration of how they can be usefully deployed to understand neural mechanism and cognitive function. That is, if deep learning is the answer, what are the questions that neuroscientists should ultimately be asking?

One strength of the deep learning framework is its generality: it offers a unified vision for studying computation across functions, species and brain regions. However, it has yet to provide a concrete road map for systems neuroscience research[23,32–34]. If neural computation emerges uncontrollably through blind, unconstrained optimization, how can neuroscientists formulate new, empirically testable hypotheses about brain function? Such hypotheses are argued to take the form of design choices about learning rules, regularization principles or architectural constraints in deep networks[33]. There is some evidence that more judicious design choices for deep networks may permit a closer match to biology[29]. For example, adding recurrent connections improves the fit to neural data[18], especially for those natural images that are harder to classify and at later poststimulus time points[17], whereas including a biologically plausible front end (a 'retina net') encourages the formation of realistic coding properties, including cell types typically found in the thalamus[52]. In general, however, we lack overall guiding principles for making such design choices. In machine learning research, networks are rarely built with biological plausibility in mind, and so there is relatively little prior guidance in how they might be used to model neural systems. Moreover, understanding the mapping from design to performance in deep networks is challenging, which is presumably why AI has a relatively poor track record in conducting interpretable or overtly hypothesis-driven research, preferring to focus instead on whether a system works rather than why it works[53].

At worst, the deep learning framework seems to face neuroscience with an existential challenge. The research programme asks researchers to document how different architectures or algorithms can encourage deep networks to form semantically meaningful representations or exhibit complex behaviours, as humans and other animals do. This endeavour sounds suspiciously similar to contemporary AI research itself. The deep learning framework seems to break with a long tradition of searching for explanations of neural computation in biological brains. Rather, it seems to propose sweeping away existing knowledge about how specific classes of computation underpin behaviour, merging the goals of theoretical neuroscience with those of contemporary AI research.

We recognize the promise of the deep learning framework and are excited about


a
Learning phase
Early
Middle
Late
Error
Time
RSA
Early
Middle
Late
IT
CNN
Dissimilar
Similar
Item
b
One-versus-all object discriminability
Primate zone
1.0
0.0
Human consistency
Pixels
V1
AlexNet
NYU
VGG
GoogleNet
ResNet-50
Inception-v3
ResNet-50 performance
Train
Test
Super-human
Chance
c
Activity
Neuron
1
2
3
Image
MEI
Predicted activity
1
Activity
Unit
1
2
3
Image

◀ Fig. 2 | **Emerging methods for comparing deep learning and the brain. a** | Comparing representational change throughout learning. Top: over the course of learning and development, behaviour may systematically improve (schematized here as reduction in error on a task). Bottom: experiments can track how neural representations change during learning (schematized as representational similarity analysis (RSA) matrices in inferior temporal cortex (IT) at three time points; top row), and whether these changes are predicted by deep networks trained using specific learning rules (schematized as RSA matrices of a convolutional neural network (CNN); bottom row). Comparing learning trajectories can help assess whether the learning procedure, rather than only the final representations, in deep neural networks is similar to that in the primate brain. **b** | Finer-grained comparisons of behaviour. Top: measuring discriminability of one image against distractor objects isolates behavioural variance that is specifically image-driven but not predicted by the object. Patterns of confusion among individual images are shared by humans (*y* axis) and macaques (primate zone) but not deep networks. Light blue bars show the human-performance consistency of models based on low-level visual representations, while dark blue bars show human-performance consistency of publicly available deep neural networks (VGG is a model developed by the Visual Geometry Group at the University of Oxford, NYU is a model developed at New York University)[11]. Bottom: classification performance of ResNet-50 trained from scratch on ImageNet (a database of images of objects) is close to perfect when it is trained and tested on standard colour images (left) and when it is trained and tested on images with additive uniform noise (middle). However, when it is trained on images with salt-and-pepper noise and tested on images with uniform noise (right), performance is at chance even though the noise types do not seem different to human observers[9]. This is one way in which humans generalize better than current deep networks. **c** | Causal tests of deep learning models. One approach, illustrated schematically here, uses 'closed-loop' experimental designs to test the predictive power of deep networks[57,58,149]. In one study[149], natural images were presented to mice while evoked neural activity (dark blue traces in the top-right panel) was recorded, and a deep neural network was trained to predict this activity (the light blue trace illustrates correspondence between unit 1 and neuron 1 in the bottom-right panel). The deep network was then used to compute a maximally exciting input image (MEI) that strongly activates that particular neuron in the model (peach pathway). This MEI is then shown to the mouse, and the resulting neural response is measured (peach part of the top-right panel). If the deep network captures the mapping from pixels to neural response, the MEI should also strongly excite the biological neuron (neuron 1). Part **b** adapted from REF.[11], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Part **b** adapted with permission from REF.[9], Neural Information Processing Systems. Part **c** adapted from REF.[149], Springer Nature Limited. Part **c** left insert (chairs) credit: Wolfgang Flamisch/Getty. Part **c** middle insert (mountain) credit: KDP/Getty.

the new possibilities offered by neural network models as theories of neural computation. We believe the strongest version of this framework will build on existing neural theories and maintain a pointed focus on explaining computation in biological brains. In other words, we hope that deep learning will provide not just a framework for neuroscience research but also a set of explicit hypotheses about behaviour, learning dynamics and neural representation in biological networks.

Deep networks as neural models

The deep learning framework is built on the proposal that neural networks learn representations and computations that resemble those in biological brains (FIG. 2). However, it is possible that the equivalence between deep networks and animal brains has been overstated[22]. Indeed, comparing the multivariate representations in brains and neural models is fraught with statistical challenges[54]. Currently, one popular approach is to learn a linear mapping from network units to neurons, and to evaluate the predictive validity of the resulting regression model in a held-out dataset. If this approach is adopted for image classification, the highest-performing deep networks can explain an impressive 60% of the variance in neuronal responses in the primate inferior temporal cortex. However, neural networks that perform substantially worse at image classification explain just 5% less[55]. Indeed, the difference in the accuracy of predictions of blood oxygen level-dependent signals between trained networks and untrained networks (that is, those with random weights) is quite small — on the order of 5–10% accuracy difference for most visual regions[19]. It is often forgotten that landmark studies on which claims of equivalent representations in deep networks and the brain are based actually used deep networks that were not trained with gradient descent[13]. It is not fully clear, then, whether existing evidence strongly separates deep learning from the more generic notion of computation in a densely wired multilayer network. Thus, an important goal for future research is going to be to more rigorously and systematically assess the status of the claim that deep networks and biological brains learn in similar ways, for example by measuring and comparing changes in representations over learning (FIG. 2a).

Testing whether neural signals are a linear transform of model activations is a good start, but such a relationship could exist even if the neural patterns in brains and neural networks differ wildly in terms of sparsity or dimensionality. Stricter tests of shared coding are provided by methods that restrict the freedom of the mapping function, such as representational similarity analysis[56], in which representations are characterized by the distance between population activity vectors evoked by different inputs. Representational similarity analysis discloses a superficial resemblance between brains and networks, but this agreement may be driven principally by shared similarity structure for stimuli that are physically similar, such as faces[14]. To go beyond correlations, systems neuroscience will need to use causal assays of the predictive links between artificial and biological networks, such as harnessing network activations for novel image synthesis[57,58] (FIG. 2c). Such closed-loop experimental designs promise strong tests of the mapping between artificial and biological brains.

Another way to test the equivalence between biological and artificial systems is to study their patterns of response. This is vital because the computations in a neural system can often profitably be understood in the context of the behaviour they produce[59]. Revealingly, humans and machines make strikingly different sorts of errors in assays of object recognition. In one study, networks were prone to confuse object classes that humans and even monkeys could safely tell apart, such as dogs and guitars, and patterns of confusion among individual images were shared by humans and macaques, but not deep networks[11] (FIG. 2b, top). Similarly, humans generalize far better than deep networks to images that have been perturbed by addition of pixelated noise, or bandpass filtering[9] (FIG. 2b, bottom), and are less prone to be fooled by deliberately misleading images[7,21,22]. There is a widespread view that biological vision exhibits a robustness that is currently lacking in supervised deep neural networks.

More generally, animal behaviour is richly structured, in theory permitting researchers to make systematic comparisons with machine performance[60]. For example, animal decisions are subject to stereotyped biases, but also irreducibly noisy[61]; animals are flexible but forgetful, behaving as if memory and control systems were capacity limited[62]; and the rate and effectiveness of information acquisition depend strongly on the structure and order of study materials[63]. Mature theories of biological brain function should be able account for these phenomena, and we hope that future deep network models will be held to this standard.

Thus far, we have argued that neural networks, and in particular modern tools from deep learning, have great potential to shape our theories of neural computation. However, we have offered two reasons to be cautious. First, we should take care not to overstate the extent to which existing experimental comparisons between deep networks and biological systems endorse deep learning as a framework for biology. Second, if we wish to use deep learning as a framework for neuroscience, it is important to be clear about what new research questions it allows us to ask. If we wish to adjust learning rules or architectures to model biological systems, where do we begin? What empirical phenomena might deep networks predict that conventional models from classical neuroscience might not? Which theories can we validate or falsify? In what follows, we take steps towards answering these questions.

Learning rules for perception

Perception provides a key opportunity to test several planks of the deep learning hypothesis. For instance, psychologists and neuroscientists have long debated the extent to which perceptual representations are prespecified by evolution or learned via experience[64]; as an example, whether primate face representations are innate or acquired remains controversial[63,65]. The deep learning hypothesis reframes this debate by asking whether neural codes could emerge from a learning principle applied to a relatively generic architecture and starting point. One strong candidate is supervised learning with gradient descent, in which representations are sculpted by feedback about the label, name or category associated with a sensory input[29]. As detailed earlier herein, supervised models have been a major focus of comparisons between deep networks and biology[29]. However, a long tradition in neuroscience emphasizes unsupervised principles such as Hebbian learning, or relatedly that representations are formed by a pressure to accurately predict the spatially or temporally local environment under an efficiency constraint[66–68]. Indeed, recent deep generative models show a remarkable ability to disentangle complex, high-dimensional signals into their latent factors under this latter self-supervised objective[69,70]. By contrast, a successful AI model that has yet to impact neuroscience proposes instead that representation formation is driven by the need to accurately predict the motivational value of experience[5]. It remains to be seen whether some combination of these more complex learning mechanisms can account for the full diversity of perceptual neural responses across modalities without building in specific domain content.

A second proposal of deep learning that is in need of testing is end-to-end learning. One way to evaluate learning rules is to assess their ability to furnish deep networks with rich representations and complex behaviours when exposed to naturalistic data. However, this approach is challenging. As noted earlier, learning may only modestly improve the match to data, suggesting that deep network representations rather than deep learning might drive much of the correspondence. Moreover, standard supervised models, such as those described earlier that are popular for explaining primate object recognition, seem to require improbable quantities of labelled data — unlike human infants, who gain sophisticated object understanding even before language is acquired. Another challenge for networks trained with gradient descent is to identify a biologically realistic implementation — that is, one where updates are local and forward and backward connectivity in the network is not required to be symmetric. Although mechanisms adopted by machine learning researchers for assigning credit to individual synapses were once thought to be biologically implausible, we now have a growing set of candidate implementations in need of empirical tests[71,72].

Given these difficulties, a more direct test of different learning principles could focus on how representations change during prolonged training. This opens the door to studies of perceptual learning that can attempt to confirm or refute these predictions[73,74]. For example, FIG. 3 shows the predictions of a neural network model trained to classify tilted gratings with gradient descent[74]. Extant neural and behavioural phenomena emerge seamlessly from the model, such as stronger sharpening of the tuning functions of the most informative neurons (FIG. 3b,c), earlier representational changes in higher cortical stages (that is, deeper layers) during training (FIG. 3d), greater proneness to transfer coarse rather than fine discrimination abilities to other non-trained stimuli (FIG. 3e) and the transfer of fine discrimination early but not late during training[73,74].

Critically, other learning principles may make qualitatively different predictions (FIG. 3f). For example, under correlational Hebbian learning, one might predict that the most active (rather than most informative) neurons would change their tuning the most. Other learning principles such as contrastive Hebbian learning[72] or predictive coding[75], which involve feedback connectivity, might predict different distributions and timings of activity changes across layers. In this way, comparisons of learning trajectories can distinguish computational principles of learning even without specifying biological implementations. This approach opens the door to a new programme of experiments focused on measuring the dynamics of representational change across cortical stages during prolonged training using macroscopic imaging techniques such as functional MRI (fMRI) or wide-field calcium imaging.

Deep learning for cognition

Deep neural networks excel at classifying complex inputs into distinct classes such as objects or words. Equally important, however, is what our brains do next: we link objects and items into diverse knowledge structures that describe our world. We know, for example, that a dog can bark and that a maple is a type of tree. Moreover, we form semantic categories from multimodal features, connecting the written and spoken name for an object with its shape, odour and texture. This conceptual knowledge of the world transcends physical appearance, interlinking diverse and even unobservable object properties (for example, that a dog has a spleen). The abstractions we acquire over the course of development form the building blocks for flexible generalization and higher-level cognition in maturity[76].

Evaluating deep learning insights beyond the realm of perceptual tasks is a key open opportunity for neuroscientists. The behaviour of humans and other animals is governed by a rich array of cognitive functions, including modular memory processes and attentional and task-level control, and neural systems for navigation, planning, mental simulation, reasoning and abstract inference. These cognitive functions are implemented in a regionally specialized brain, in which a patchwork of subcortical and allocortical structures interconnects with granular and infragranular cortical zones, each housing unique cell types and circuits. If we are committed to deploying deep learning models as theories for biology, we need to take seriously the question of how such elaborate structure in cognition and behaviour emerges via optimization. How do humans learn abstract representations, divorced from physical object properties? How do we assemble knowledge into relational structures such as trees, rings and grids? How do we

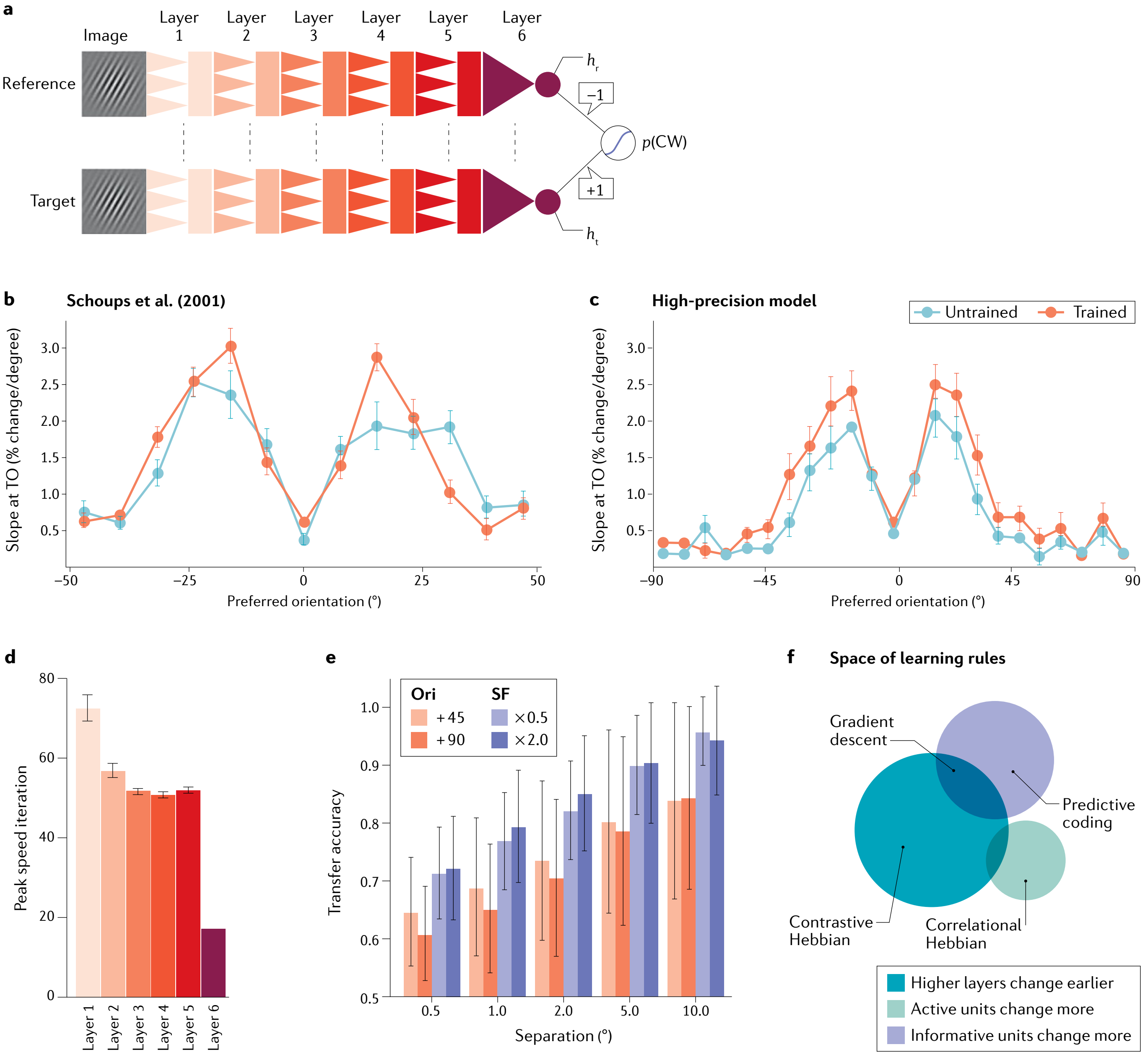


Fig. 3 | **Testing principles of learning using perceptual learning paradigms. a** | A deep network model of perceptual learning[74]. Clockwise-oriented or anticlockwise-oriented visual inputs flow through layers of weights to an output layer that reports the direction of rotation. h_t and h_r denote the last hidden unit for the target and reference images, respectively; p(CW) is the probability that the target image is clockwise with respect to the reference. **b** | Tuning curve slope changes due to learning measured in primate primary visual cortex (V1)[150]. **c** | Tuning curve slope changes due to gradient descent learning in the model illustrated in part **a**[74]. **d** | Timing of peak synaptic changes in each layer. The weights at higher layers change earlier[74]. **e** | Behavioural performance transfer to different orientations (Ori) and spatial frequencies (SF) after training on tasks of different angular separation[74]. **f** | A schematized conceptual space of learning rules, in which specific learning rules are points. Experimental observations may span large regions of this space (coloured circles) and thus could be theoretically consistent with multiple learning rules. Intersecting many constraints can begin to narrow the set of candidate learning algorithms. TO, trained orientation. Parts **a**, **c**, **d** and **e** adapted from REF.[74], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). Part **b** adapted from REF.[150], Springer Nature Limited.

compose new behaviours from existing subcomponents? How do we rapidly acquire and generalize new memories? These are important questions for AI researchers as well, and indeed, some have expressed a hope that machine learning will soon offer more powerful models in which higher cognitive functions emerge naturally via a ‘blind search’ process, allowing neuroscientists to sidestep the problem of modelling them explicitly[32]. Indeed, recent advances in AI research have followed the successful fusion of deep learning with other methods, such as reinforcement learning[5], content-addressable memory[77,78] or Monte Carlo tree search[4], demonstrating a proof of concept for end-to-end learning in complex cognitive architectures.

Neuroscientists can harness their familiar experimental toolkit to study cognition using deep networks, pressing towards more complex behaviours and exposing current limitations of the deep learning hypothesis. One potentially fruitful approach is to identify specific problems or tasks in which human performance and network performance are qualitatively different. For example, if networks struggle with one problem class whereas humans do not, one can use techniques from psychology and neuroscience to identify how humans solve the problem, and then ask how the learning rule or architecture may be varied to accommodate the same behaviour in deep networks. Experimental interrogation of interacting brain systems can help to identify constraints on network submodules that generate forms of information processing that resemble memory, attention or reasoning systems in biological agents. Thus, the research programme should build upon the work of past decades, in which neuroscientists have experimentally dissected cognitive systems, in many cases providing a detailed, computationally grounded account of their function. For example, we understand a great deal about the navigation system in the rodent medial temporal lobe (MTL)[79], the motor system in songbirds[80] and the saccadic system in the macaque[81]. With this foundation, deep learning approaches can help explain key computational trade-offs and underlying reasons for the specialized, interacting subsystems evident in biology.

Abstraction and generalization

Deep networks excel when data are abundant and training is exhaustive. However, they struggle to extrapolate this knowledge to new environments populated by previously unseen features and objects. Humans, by contrast, seem to generalize effectively[28]. For example, most people can navigate a foreign city where the language, coinage and customs are unfamiliar, because they understand concepts such as 'greeting', 'taxi' and 'map'. A popular view is that deep networks fail to transfer knowledge because they do not form neural codes that abstract over physically dissimilar domains. Building deep networks that can generalize in this way would be a major milestone for machine learning. In parallel, this difference provides an incentive for neuroscientists to study how biological brains encode, compose and generalize abstract knowledge[82–84].

Unfortunately, key methodological challenges arise for neuroscientists seeking to address this question. First, it is unknown whether experimental animals such as rodents and macaques (or even our closer primate cousins)[85] have evolved neural mechanisms that permit the strong, flexible transfer of knowledge characteristic of human intelligence. It is thus unclear whether invasive tools for recording and interference (such as electrophysiology or optogenetics) can be used to study generalization and transfer in animals. Moreover, to study human abstraction, we are obliged to use macroscopic imaging methods such as fMRI, magnetoencephalography and electroencephalography that are less well suited to revealing how computation unfolds in neural circuits. Nevertheless, inventive new ways of using these tools are being developed, allowing researchers to probe replay[86–88], changes in excitatory–inhibitory balance[89,90] and hexagonal (grid) coding[91,92] in human brain signals. Second, humans (and other animals) usually enter the laboratory with rich past experiences that sculpt the ways in which they learn. This complicates direct comparisons between humans and neural networks, because it is difficult to imbue artificial systems with equivalent priors, or to eliminate human priors using wholly novel stimuli. Third, humans and neural networks learn over very different timescales. For example, deep reinforcement learning systems exceed human performance on Atari video games but require many times more training than a human player to acquire this proficiency[93].

In an end-to-end learning system, abstract representations need to be grounded in experience. One possibility is that lifelong exposure to huge volumes of sensory data might allow strong invariances to emerge naturally via either supervised or unsupervised learning. There is evidence that cells in the MTL, which sits at the apex of the primate ventral stream, develop physically invariant coding properties. For example, in humans, 'concept' cells encode famous individuals or landmarks, irrespective of whether they are denoted by pictures or words[94]. Echoes of this coding scheme in which MTL coding is tied more tightly to allocentric space can be seen in other animals. For example, in rodents, hippocampal place cells encode locations in a way that is invariant to viewpoint and heading direction[95], and in primates, 'schema' cells activate across environments in a way that allows for generalization over common spatial configurations[96]. These neural codes for high-level concepts can form when different features, objects or locations are repeatedly associated in space or time, for example through Hebbian learning[97]. Indeed, fMRI studies of statistical learning have revealed that neural similarities (such as multivoxel pattern overlap) in the MTL recapitulate association strengths for pairs, lines, maps or hierarchies of stimuli[94,98–104]. Moreover, in a bandit task, the entorhinal cortex is one brain region where a consistent mapping exists between neural patterns and the covariance among stimuli and rewards, irrespective of the physical images involved[105]. Stitching together multiple patterns of association, and learning their structures, could enable animals to learn a comprehensive model of the world that can be used for navigation, inference and planning[106].

In parallel with this growing emphasis on the virtues of model-based computation in neuroscience, machine learning researchers are building powerful deep generative models that are capable of disentangling the world into its latent factors, and recomposing these to construct realistic synthetic 3D images[70,107,108]. However, to date, wiring these generative models up with control systems to build intelligent agents has proved challenging, despite some promising efforts[78]. Indeed, AI researchers have struggled to build model-based systems that can hold their own against model-free agents in benchmark problems such as Atari games[109]. It is paradoxical that this is occurring against a rich backdrop of neuroscience research that emphasizes the virtues of model-based inference. Neuroscientists have even begun to unravel how seemingly idiosyncratic coding properties in the MTL and other structures may be hallmarks of a normative scheme for computationally efficient planning and inference[110,111]. For example, grid cell codes in the medial entorhinal cortex and elsewhere may be signatures of a neural code that has learned the geometry by which space itself is structured, potentially supporting transfer learning for navigation[111]. There are even hints that this coding scheme may apply to non-spatial as well as spatial domains[92], potentially laying the foundations for a theory of higher-order human reasoning[112]. Although machine learning researchers have noted that lattice-like codes may emerge when deep reinforcement learning systems are trained to navigate[113,114], they have yet to build on these insights for building stronger AI. More generally, understanding how to simulate biologically plausible model-based computations in a way that is useful to machine learning researchers is a potentially rich intellectual seam that neuroscientists are only just beginning to exploit.

Resource allocation in learning

Humans and other animals continue to learn across their lifespan. This 'continual' learning might allow a human to acquire a second language, a monkey to adopt a new social role or a rodent to navigate in a novel environment. This is in stark contrast to most current AI systems, which lack the flexibility to acquire new behaviours once they have achieved convergence on an initial task. Building machines that can learn continually, as humans and other animals do, is proving one of the thorniest challenges in contemporary machine learning research[115]. Fortunately, however, this question has opened up new avenues for neuroscience research focused on how biology may have solved continual learning[8,116].

It has long been noted that, in neural networks, learning pursuant to an initial task A is often overwritten during subsequent training on task B (known as 'catastrophic interference')[117]. This occurs because a parameterization that solves task A is not guaranteed to solve any other task, and so during training on task B, gradient descent drives network weights away from the local minimum (that is, a setting that specifies a local optimum for behaviour) for task A. It occurs even when the network has sufficient capacity to perform both tasks, because simultaneous (or 'interleaved') training allows the discovery of a setting that jointly solves tasks A and B. In humans, new learning can sometimes degrade extant performance, for example when memorizing associate pairs A–C after having encoded pairs A–B, but in general interference effects are far less dramatic than for neural networks[118].

One popular model suggests that mammals have evolved to solve continual learning by using complementary learning systems in the hippocampus and neocortex[116,119,120]. Unlike the cortex, the hippocampus can rapidly learn sparse (or 'pattern-separated') representations of specific experiences, often called 'episodic memories'[121], and these memories are replayed offline during periods of rest or sleep[122]. Hippocampal replay provides an opportunity for virtual interleaving of past and present experiences, potentially allowing memories to be gradually consolidated into neocortical circuits in a way that circumvents the problem of catastrophic interference. This theory is supported by a wealth of evidence, including the finding that hippocampal damage leads to a gradient of retrograde amnesia[123], and reports of double dissociations between instance-based memory (or 'recollection') in the hippocampus and summaries of past experience (or 'familiarity') in the neocortex[124]. In more recent years, artificial replay of past experiences has emerged as a crucial factor that enables deep networks to exhibit strong performance in temporally correlated environments[125], including deep reinforcement learning agents for dynamic video games[5]. Pleasingly, this has allowed theorists to draw a link between computational solutions to continual learning in biological intelligence and AI[126]. Adaptations of the complementary learning system framework allow it to account for seemingly contradictory phenomena, such as the involvement of MTL structures in rapid statistical learning[116].

Although evidence that offline replay may be important for memory consolidation has grown, the problem of continual learning has prompted new questions for neuroscientists. Is biological learning actively partitioned so as to avoid catastrophic interference? Unlike neural networks, animals do not always benefit from interleaved study conditions (imagine learning the violin and the cello at once). For example, humans who have been trained in a blocked regimen to classify naturalistic stimuli (trees) according to two orthogonal boundaries (their 'leafiness' versus their 'branchiness') perform better on a later interleaved test compared with those who experienced the same conditions at training and test[8]. Other evidence from human category learning implies that human knowledge may be actively partitioned by time and context[127,128]. Indeed, promising solutions to continual learning in the machine learning literature rely on the identification of weight subspaces in which new learning is least likely to cause retrospective interference, for example by 'freezing' synapses that are more likely to participate in extant tasks[129,130]. These tools are more effective when coupled with a gating process that overtly earmarks neural subspaces for new learning, in a way that resembles top-down attention in the primate neocortex[131,132]. Another intriguing possibility is that unsupervised processes facilitate continual learning in biological systems by clustering neural representations according to their context. Hebbian learning might encourage the formation of orthogonal neural codes for different temporal contexts[133], which in turn enables tasks to be learned in different neural subspaces[132]. The curious phenomenon of 'representational drift' (in which neural codes meander unpredictably over time)[134] might reflect the allocation of information to different neural circuits in distinct contexts, enabling task knowledge to be partitioned in a way that minimizes interference[135].

A more general question concerning resource allocation is how biological systems have evolved both to minimize negative transfer (interference) and maximize positive transfer (generalization) among tasks. One fascinating theoretical perspective argues that the capacity limits that are inherent in biological control processes are a response to this conundrum[136]. Using simulations involving deep networks, Musslick et al. show that shared and separate task representations have mixed costs and benefits, with shared codes enabling generalization between tasks at the risk of interference between tasks. They suggest that the brain has found a solution by promoting shared neural codes, which in turn enables strong transfer, but deploying control processes to gate out irrelevant tasks that might provoke interference. They suggest that this answers the question of why, despite a brain that comprises billions of neurons and trillions of connections, humans struggle with multitasking problems such as typing a line of computer code while answering a verbal question[136].

Understanding deep networks

To fully realize the promise of deep neural networks as scientific theories of brain function, we need to understand how they work. Unfortunately, the computations performed by deep networks are enmeshed in millions of trainable parameters, such that they have been dubbed 'black boxes'. Despite this complexity, however, in neural networks we can access every synaptic weight and unit activation over the course of learning, a feat that remains impossible in animal models. These considerations raise thorny questions concerning the utility of deep networks as neural models, and more generally, what it means to 'understand' a neural process via a computational model[34].

Thus far, many neuroscientists drawing on the deep learning toolkit have preferred to use simulations of off-the-shelf black box deep networks as neural models[29]. However, collaborations between theoretical neuroscientists, physicists and computer scientists have paved the way for a new approach that uses idealized neural network models to understand the mathematical principles by which they learn[137], and deploys the results to predict or explain phenomena in psychology or neuroscience[138]. For this endeavour to be tractable, deep network models must be simplified (FIG. 4), for example by using linear activation functions ('deep linear' networks)[139] (FIG. 4a–c) or specially

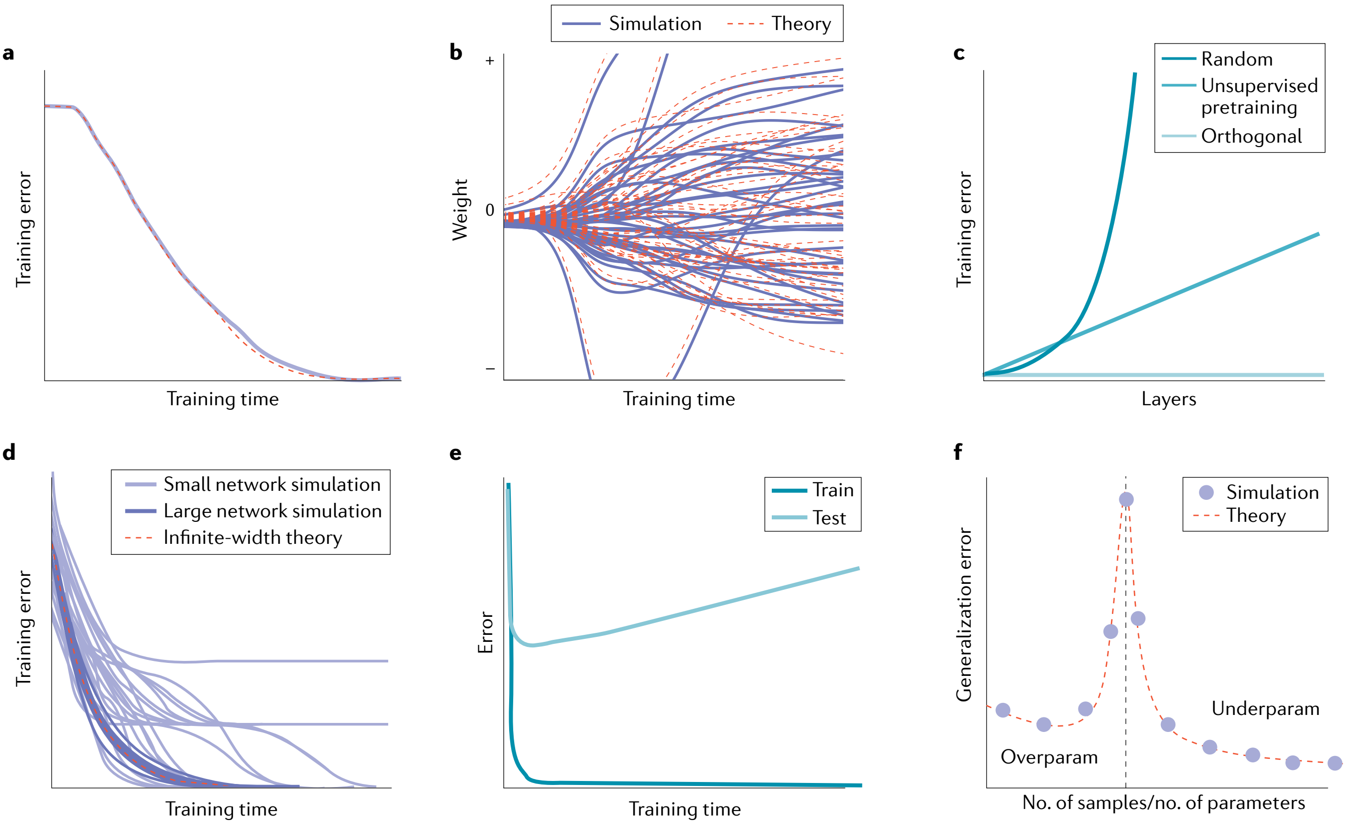


Fig. 4 | **Understanding deep networks using idealized models.** Although mathematical insight into general non-linear deep networks remains challenging, there are a growing number of well-understood simplified settings. **a**,**b** | The error-corrective learning process in deep neural networks is often simulated on a computer and can exhibit complex training error dynamics (part **a**) and complex synaptic weight dynamics (part **b**) as schematized here (solid curves). By simplifying the neural non-linearity, deep linear networks permit exact analytical solutions for training error dynamics (part **a**) and weight dynamics (part **b**) from certain initializations, plotted as dotted curves[139]. These solutions explicitly describe the trajectory of every weight over the course of training, removing the need to simulate these networks and directly revealing the impact of dataset statistics on learning dynamics. **c** | Analytical solutions have shed light on various phenomena, including how training speed in deep linear networks depends on network initialization. As schematized here, deep linear networks starting with small random weights train exponentially slowly as depth increases, those with unsupervised layerwise pretraining train linearly and training speed in networks with large orthogonal initializations is depth independent, consistent with empirical findings from simulations of non-linear networks[137,139]. **d** | Learning dynamics can simplify in very large 'wide' networks with many non-linear neurons. Schematic of training error of non-linear networks of different sizes trained on the same task from different random initial conditions. Simulations of small networks with few neurons often exhibit complex trajectories that end at local minima with non-zero error (light purple). By contrast, simulations of large networks with many neurons reliably find a zero-error solution and take similar trajectories (dark purple). Remarkably, as the number of neurons tends to infinity in a specific initialization regime, the trajectory can be described analytically[142,143] (dashed red line). **e**,**f** | Tractable settings can also arise by placing assumptions on how data are generated. In one influential approach, a 'teacher' neural network labels data for a 'student' neural network. This setting allows analytical description of both training (blue) and testing (light blue) error, schematized in part **e**, and permits analysis of the overtraining phenomenon[140,141,144,145]. As schematized in part **f**, the student–teacher setting enables analytical predictions (dashed red line) for the generalization error in the 'high-dimensional' regime, where data are scarce relative to the number of weights. These predictions closely match the performance of simulated large networks (purple dots), and explain why generalization error peaks at the transition from overparameterization (overparam) to underparameterization (underparam)[141,145,151].

structured environments[140,141]. Often the behaviour of deep networks becomes simpler in 'limit' cases, such as when the number of neurons per layer diverges towards infinity (infinite width limit)[142,143] (FIG. 4d) or when the number of data samples and model parameters both diverge towards infinity but their ratio is finite (the high-dimensional limit)[144,145] (FIG. 4e,f). Paradoxically, infinite-size networks can be more interpretable than those with fewer units, because their learning trajectory is stabler and not prone to being waylaid by bad local minima in the loss landscape, leading to suboptimal outcomes[141,144,146] (FIG. 4d). Leveraging these simplifying assumptions has allowed researchers to derive exact solutions for the learning trajectories that every single synapse will follow in certain networks[139,142,143] (FIG. 4a,b,d). These network idealizations have generated mathematical insight into perplexing questions about network behaviour, including why deep networks are often slower to train[138] (FIG. 4c), why an initial epoch of layer-by-layer statistical learning reminiscent of critical period plasticity ('unsupervised pretraining') can accelerate future learning with gradient descent[139] (FIG. 4c) and why generalization to unseen data suffers at the transition to overparameterization[144–146] (FIG. 4e,f). This work challenges the notion that deep neural networks are black boxes and promises interpretable neural network models of biological phenomena.

Recently, this approach has been applied to the study of semantic cognition[138] (FIG. 5). During development, children transition through quasi-discrete stages in which they rapidly acquire new categories or concepts. Their learning is also highly structured: for example, semantic knowledge is progressively differentiated, as children pick up on broader hierarchical distinctions ('animal' versus 'plant') before finer distinctions ('rose' versus 'daisy'), and displays stereotyped errors (such as thinking that worms have bones)[147]. Deep networks trained on richly structured data (FIG. 5a) are known to exhibit these phenomena[148], but only recently has it been shown: that stage-like transitions arise due to so-called saddle points in the error surface (FIG. 5c), that progressive differentiation arises from the way the singular values of the input–output correlations drive learning over time (FIG. 5a–d) and that semantic illusions arise from pressure to sacrifice accuracy on exceptions to meet the global supervised objective[138] (FIG. 5e). Moreover, these phenomena can be shown to be a consequence of depth itself, arising in deep linear networks but not shallow networks (FIG. 5c,e), even though the two classes of model converge to identical terminal solutions. This highlights the importance for neuroscientists of studying learning dynamics — that is, the trajectory that learning takes — rather than simply examining representations in networks that have converged.

One potential concern is that insights acquired in this way might not scale, because models are idealizations that eschew the messy complexity of state-of-the art deep networks and make assumptions that are false for biology (such as linear transduction, or layers of infinite width). However, we argue that neural theory is well served by analytical formulations of complex phenomena that give rise to specific, falsifiable predictions for neural circuits and systems. We hope that neuroscientists will incorporate reductions of deep network models into their canonical set of neural theories, rather than only seeking correspondences between brains and fully fledged deep learning systems that offer little hope of being understood.

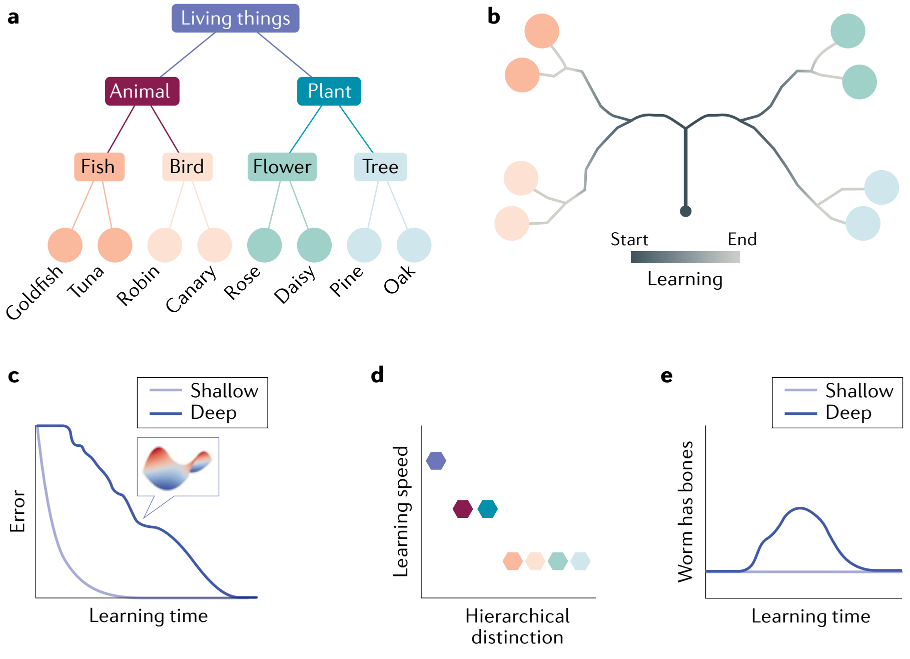


Fig. 5 | **Developmental trajectories in deep linear neural networks. a** | An idealized hierarchical environment. Items (leaf nodes) possess many properties, such as 'can fly' or 'has roots'. Nearby items in the tree are more likely to share properties. **b** | Two-dimensional embedding of internal representations for each item over learning in a deep linear network trained to output each item's properties. The network exhibits progressive differentiation, passing through a series of stages in which higher hierarchical distinctions are learned before lower distinctions. **c** | As schematized here, only deep networks exhibit quasi-stage-like transitions in learning, which arise from saddle points in the error surface[138]. **d** | For a class of hierarchies, learning speed decreases as a function of hierarchical level (indicated by colour, as in part **a**), and the network exhibits progressive differentiation starting with the broadest distinction. **e** | Deep but not shallow networks can make transient errors on specific items and properties (such as asserting that 'worms have bones') during learning reminiscent of human semantic development. Parts **a**, **b**, **d** and **e** adapted with permission from REF.[138], National Academy of Sciences, USA. Insert in part **c** adapted with permission from REF.[137], Annual Reviews.

Conclusions

Deep learning models have much to offer neuroscience. Most exciting is the potential to go beyond handcrafting of function, and to understand how computation emerges from experience. Neuroscientists have recognized this opportunity, but its exploitation has only just begun. In this Perspective, we have tried to offer a road map for researchers wishing to use deep networks as neural theories. Our principal exhortation for neuroscientists is to use deep networks as predictive models that make falsifiable predictions, and to use model idealization methods to provide genuine understanding of how and why they might capture biological phenomena. We caution against using increasingly complex models and simulations that outpace our conceptual insight, and discourage the blind search for correspondences in neural codes formed by biological and artificial systems. Instead, we hope that neuroscientists will build models that explain human behaviour, learning dynamics and neural coding in rich and fruitful ways, but without losing the interpretability inherent to classical neural models.

Andrew Saxe ✉, *Stephanie Nelli* ✉
and Christopher Summerfield ✉

Department of Experimental Psychology,
University of Oxford, Oxford, UK.

✉*e-mail: andrew.saxe@psy.ox.ac.uk; stephanie.nelli@psy.ox.ac.uk; christopher.summerfield@psy.ox.ac.uk*

https://doi.org/10.1038/s41583-020-00395-8

Published online: 16 November 2020

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Krizhevsky, A., Hinton, G. E. & Sutskever, I. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.* **25**, 1106–1114 (2012).
3. Eslami, S. M. A. et al. Neural scene representation and rendering. *Science* **360**, 1204–1210 (2018).
4. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
5. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
6. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron* **95**, 245–258 (2017).
7. Golan, T., Raju, P. C. & Kriegeskorte, N. Controversial stimuli: pitting neural networks against each other as models of human recognition. Preprint at *arXiv* https://arxiv.org/abs/1911.09288 (2020).
8. Flesch, T., Balaguer, J., Dekker, R., Nili, H. & Summerfield, C. Comparing continual task learning in minds and machines. *Proc. Natl Acad. Sci. USA* **115**, E10313–E10322 (2018).
9. Geirhos, R. et al. Generalisation in humans and deep neural networks. *NeurIPS Proc.* (2018).
10. Zhou, Z. & Firestone, C. Humans can decipher adversarial images. *Nat. Commun.* **10**, 1334 (2019).
11. Rajalingham, R. et al. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* **38**, 7255–7269 (2018).
12. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).

13. Yamins, D. L. K. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
14. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
15. Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V. & McDermott, J. H. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* **98**, 630–644.e16 (2018).
16. 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).
17. 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 behavior. *Nat. Neurosci.* **22**, 974–983 (2019).
18. 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).
19. 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).
20. Elsayed, G. F. et al. Adversarial examples that fool both computer vision and time-limited humans. *NeurIPS Proc.* (2018).
21. Ullman, S., Assif, L., Fetaya, E. & Harari, D. Atoms of recognition in human and computer vision. *Proc. Natl Acad. Sci. USA* **113**, 2744–2749 (2016).
22. Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M. & Tolias, A. S. Engineering a less artificial intelligence. *Neuron* **103**, 967–979 (2019).
23. Marblestone, A. H., Wayne, G. & Kording, K. P. Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.* **10**, 1–61 (2016).
24. Kell, A. J. & McDermott, J. H. Deep neural network models of sensory systems: windows onto the role of task constraints. *Curr. Opin. Neurobiol.* **55**, 121–132 (2019).
25. Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015).
26. Bowers, J. S. Parallel distributed processing theory in the age of deep networks. *Trends Cogn. Sci.* **21**, 950–961 (2017).
27. Cichy, R. M. & Kaiser, D. Deep neural networks as scientific models. *Trends Cogn. Sci.* **23**, 305–317 (2019).
28. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci*. **40**, e253 (2017).
29. Lindsay, G. W. Convolutional neural networks as a model of the visual system: past, present, and future. *J. Cogn. Neurosci.* https://doi.org/10.1162/jocn_a_01544 (2020).
30. Zador, A. M. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat. Commun*. **10**, 3770 (2019).
31. Rogers, T. T. & Mcclelland, J. L. Parallel distributed processing at 25: further explorations in the microstructure of cognition. *Cogn. Sci.* **38**, 1024–1077 (2014).
32. Hasson, U., Nastase, S. A. & Goldstein, A. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron* **105**, 416–434 (2020).
33. Richards, B. A. et al. A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).
34. Lillicrap, T. P. & Kording, K. P. What does it mean to understand a neural network? Preprint at *arXiv* https://arxiv.org/abs/1907.06374 (2019).
35. Saxe, A., Bhand, M., Mudur, R., Suresh, B. & Ng, A. Y. Unsupervised learning models of primary cortical receptive fields and receptive field plasticity. *Adv. Neural Inform. Process. Syst.* **25**, 1971–1979 (2011).
36. Stevenson, I. H. & Kording, K. P. How advances in neural recording affect data analysis. *Nat. Neurosci.* **14**, 139–142 (2011).
37. Saxena, S. & Cunningham, J. P. Towards the neural population doctrine. *Curr. Opin. Neurobiol.* **55**, 103–111 (2019).
38. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).
39. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
40. Johnston, W. J., Palmer, S. E. & Freedman, D. J. Nonlinear mixed selectivity supports reliable neural computation. *PLoS Comput. Biol.* **16**, e1007544 (2020).
41. Raposo, D., Kaufman, M. T. & Churchland, A. K. A category-free neural population supports evolving demands during decision-making. *Nat. Neurosci.* **17**, 1784–1792 (2014).
42. Higgins, I. et al. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal neurons. Preprint at *arXiv* https://arxiv.org/abs/2006.14304 (2020).
43. Park, I. M., Meister, M. L. R., Huk, A. C. & Pillow, J. W. Encoding and decoding in parietal cortex during sensorimotor decision-making. *Nat. Neurosci.* **17**, 1395–1403 (2014).
44. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
45. Chaisangmongkon, W., Swaminathan, S. K., Freedman, D. J. & Wang, X.-J. Computing by robust transience: how the fronto-parietal network performs sequential, category-based decisions. *Neuron* **93**, 1504–1517.e4 (2017).
46. Engel, T. A., Chaisangmongkon, W., Freedman, D. J. & Wang, X. J. Choice-correlated activity fluctuations underlie learning of neuronal category representation. *Nat. Commun.* **6**, 6454 (2015).
47. Remington, E. D., Egger, S. W., Narain, D., Wang, J. & Jazayeri, M. A dynamical systems perspective on flexible motor timing. *Trends Cogn. Sci.* **22**, 938–952 (2018).
48. Remington, E. D., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible sensorimotor computations through rapid reconfiguration of cortical dynamics. *Neuron* **98**, 1005–1019.e5 (2018).
49. 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).
50. Masse, N. Y., Rosen, M. C. & Freedman, D. J. Reevaluating the role of persistent neural activity in short-term memory. *Trends Cogn. Sci.* **24**, 242–258 (2020).
51. Masse, N. Y., Yang, G. R., Song, H. F., Wang, X.-J. & Freedman, D. J. Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nat. Neurosci.* **22**, 1159–1167 (2019).
52. Lindsey, J., Ocko, S. A., Ganguli, S. & Deny, S. A unified theory of early visual representations from retina to cortex through anatomically constrained deep CNNs. Preprint at *bioRxiv* https://doi.org/10.1101/511535 (2019).
53. Rahwan, I. et al. Machine behaviour. *Nature* **568**, 477–486 (2019).
54. Thompson, J. A. F., Bengio, Y., Formisano, E. & Schönwiesner, M. How can deep learning advance computational modeling of sensory information processing? Preprint at *arXiv* https://arxiv.org/abs/1810.08651 (2018).
55. 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 (2018).
56. Kriegeskorte, N. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
57. Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).
58. 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).
59. Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D. Neuroscience needs behavior: correcting a reductionist bias. *Neuron* **93**, 480–490 (2017).
60. Gomez-Marin, A. & Ghazanfar, A. A. The life of behavior. *Neuron* **104**, 25–36 (2019).
61. Rich, A. S. & Gureckis, T. M. Lessons for artificial intelligence from the study of natural stupidity. *Nat. Mach. Intell.* **1**, 174–180 (2019).
62. Shenhav, A., Botvinick, M. M. & Cohen, J. D. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron* **79**, 217–240 (2013).
63. Deen, B. et al. Organization of high-level visual cortex in human infants. *Nat. Commun.* **8**, 13995 (2017).
64. Op de Beeck, H. P., Pillet, I. & Ritchie, J. B. Factors determining where category-selective areas emerge in visual cortex. *Trends Cogn. Sci.* **23**, 784–797 (2019).
65. Arcaro, M. J., Schade, P. F., Vincent, J. L., Ponce, C. R. & Livingstone, M. S. Seeing faces is necessary for face-domain formation. *Nat. Neurosci.* **20**, 1404–1412 (2017).
66. 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).
67. Simoncelli, E. P. & Olshausen, B. A. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).
68. Friston, K. J. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
69. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. Preprint at *arXiv* https://arxiv.org/abs/1312.6114 (2014).
70. Burgess, C. P. et al. MONet: unsupervised scene decomposition and representation. Preprint at *arXiv* https://arxiv.org/abs/1901.11390 (2019).
71. 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).
72. Detorakis, G., Bartley, T. & Neftci, E. Contrastive Hebbian learning with random feedback weights. *Neural Netw.* https://doi.org/10.1016/j.neunet.2019.01.008 (2019).
73. Saxe, A. *Deep Linear Networks: A Theory of Learning in the Brain and Mind.* Thesis, Stanford Univ. (2015).
74. Wenliang, L. K. & Seitz, A. R. Deep neural networks for modeling visual perceptual learning. *J. Neurosci.* **38**, 6028–6044 (2018).
75. Whittington, J. C. R. & Bogacz, R. An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity. *Neural Comput.* **29**, 1229–1262 (2017).
76. Murphy, G. L. *The Big Book of Concepts* (MIT Press, 2002).
77. Graves, A. et al. Hybrid computing using a neural network with dynamic external memory. *Nature* **538**, 471–476 (2016).
78. Wayne, G. et al. Unsupervised predictive memory in a goal-directed agent. Preprint at *arXiv* https://arxiv.org/abs/1803.10760 (2018).
79. Moser, E. I., Kropff, E. & Moser, M.-B. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.* **31**, 69–89 (2008).
80. Fee, M. S., Kozhevnikov, A. A. & Hahnloser, R. H. R. Neural mechanisms of vocal sequence generation in the songbird. *Ann. N. Y. Acad. Sci.* **1016**, 153–170 (2004).
81. Hanes, D. P. & Schall, J. D. Neural control of voluntary movement initiation. *Science* **274**, 427–430 (1996).
82. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: statistics, structure, and abstraction. *Science* **331**, 1279–1285 (2011).
83. Tervo, D. G. R., Tenenbaum, J. B. & Gershman, S. J. Toward the neural implementation of structure learning. *Curr. Opin. Neurobiol.* **37**, 99–105 (2016).
84. Behrens, T. E. J. et al. What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* **100**, 490–509 (2018).
85. Penn, D. C., Holyoak, K. J. & Povinelli, D. J. Darwin's mistake: explaining the discontinuity between human and nonhuman minds. *Behav. Brain Sci.* **31**, 109–130 (2008).
86. Schuck, N. W. & Niv, Y. Sequential replay of nonspatial task states in the human hippocampus. *Science* **364**, eaaw5181 (2019).
87. Kurth-Nelson, Z., Economides, M., Dolan, R. J. & Dayan, P. Fast sequences of non-spatial state representations in humans. *Neuron* **91**, 194–204 (2016).
88. Liu, Y., Dolan, R. J., Kurth-Nelson, Z. & Behrens, T. E. J. Human replay spontaneously reorganizes experience. *Cell* **178**, 640–652.e14 (2019).
89. Barron, H. C. et al. Unmasking latent inhibitory connections in human cortex to reveal dormant cortical memories. *Neuron* **90**, 191–203 (2016).
90. Koolschijn, R. S. et al. The hippocampus and neocortical inhibitory engrams protect against memory interference. *Neuron* **101**, 528–541.e6 (2019).

91. Doeller, C. F., Barry, C. & Burgess, N. Evidence for grid cells in a human memory network. *Nature* **463**, 657–661 (2010).
92. Constantinescu, A. O., OReilly, J. X. & Behrens, T. E. J. Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468 (2016).
93. Tsividis, P. A., Pouncy, T., Xu, J., Tenenbaum, J. B. & Gershman, S. J. Human learning in Atari (AAAI, 2017).
94. Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B. & Botvinick, M. M. Neural representations of events arise from temporal community structure. *Nat. Neurosci.* **16**, 486–492 (2013).
95. O'Keefe, J. & Dostrovsky, J. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **34**, 171–175 (1971).
96. Baraduc, P., Duhamel, J.-R. & Wirth, S. Schema cells in the macaque hippocampus. *Science* **363**, 635–639 (2019).
97. Miyashita, Y. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature* **335**, 817–820 (1988).
98. Garvert, M. M., Dolan, R. J. & Behrens, T. E. A map of abstract relational knowledge in the human hippocampal-entorhinal cortex. *eLife* **6**, e17086 (2017).
99. Schapiro, A. C., Kustner, L. V. & Turk-Browne, N. B. Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Curr. Biol.* **22**, 1622–1627 (2012).
100. Schapiro, A. C., Turk-Browne, N. B., Norman, K. A. & Botvinick, M. M. Statistical learning of temporal community structure in the hippocampus:. *Hippocampus* **26**, 3–8 (2016).
101. Schlichting, M. L., Mumford, J. A. & Preston, A. R. Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nat. Commun.* **6**, 8151 (2015).
102. Zeithamova, D., Dominick, A. L. & Preston, A. R. Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron* **75**, 168–179 (2012).
103. Park, S. A., Miller, D. S., Nili, H., Ranganath, C. & Boorman, E. D. Map making: constructing, combining, and inferring on abstract cognitive maps. *Neuron* **107**, 1226–1238.e8 (2020).
104. Kumaran, D., Banino, A., Blundell, C., Hassabis, D. & Dayan, P. Computations underlying social hierarchy learning: distinct neural mechanisms for updating and representing self-relevant information. *Neuron* **92**, 1135–1147 (2016).
105. Baram, A. B., Muller, T. H., Nili, H., Garvert, M. & Behrens, T. E. J. Entorhinal and ventromedial prefrontal cortices abstract and generalise the structure of reinforcement learning problems. Preprint at *bioRxiv* https://doi.org/10.1101/827253 (2019).
106. Dolan, R. J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).
107. Higgins, I. et al. Early visual concept learning with unsupervised deep learning. Preprint at *arXiv* https://arxiv.org/abs/1606.05579 (2016).
108. Higgins, I. et al. SCAN: learning hierarchical compositional visual concepts. Preprint at *arXiv* https://arxiv.org/abs/1707.03389 (2018).
109. Hessel, M. et al. Rainbow: combining improvements in deep reinforcement learning (AAAI, 2018).
110. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. Design principles of the hippocampal cognitive map. *Int. Conf. Neural Inform. Process. Syst.* **2**, 2528–2536 (2014).
111. Whittington, J. C. et al. The Tolman-Eichenbaum machine: unifying space and relational memory through generalisation in the hippocampal formation. Preprint at *bioRxiv* https://doi.org/10.1101/770495 (2019).
112. Bellmund, J. L. S., Gärdenfors, P., Moser, E. I. & Doeller, C. F. Navigating cognition: spatial codes for human thinking. *Science* **362**, eaat6766 (2018).
113. Cueva, C. J. & Wei, X.-X. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. Preprint at *arXiv* https://arxiv.org/abs/1803.07770 (2018).
114. Banino, A. et al. Vector-based navigation using grid-like representations in artificial agents. *Nature* **557**, 429–433 (2018).
115. Parisi, G. I., Kemker, R., Part, J. L., Kanan, C. & Wermter, S. Continual lifelong learning with neural networks: a review. *Neural Netw.* **113**, 54–71 (2019).
116. Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M. & Norman, K. A. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Phil. Trans. R. Soc. B* **372**, 20160049 (2017).
117. French, R. Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* **3**, 128–135 (1999).
118. McCloskey, M. & Cohen, N. J. Catastrophic interference in connectionist networks: the sequential learning problem. in *Psychology of Learning and Motivation* Vol. 24 109–165 (Academic, 1989).
119. McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
120. O'Reilly, R. C., Bhattacharyya, R., Howard, M. D. & Ketz, N. Complementary learning systems. *Cogn. Sci.* **38**, 1229–1248 (2014).
121. Tulving, E. Episodic memory: from mind to brain. *Annu. Rev. Psychol.* **53**, 1–25 (2002).
122. Carr, M. F., Jadhav, S. P. & Frank, L. M. Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nat. Neurosci.* **14**, 147–153 (2011).
123. Zola-Morgan, S. & Squire, L. The primate hippocampal formation: evidence for a time-limited role in memory storage. *Science* **250**, 288–290 (1990).
124. Yonelinas, A. P. The nature of recollection and familiarity: a review of 30 years of research. *J. Mem. Lang.* **46**, 441–517 (2002).
125. van de Ven, G. M. & Tolias, A. S. Generative replay with feedback connections as a general strategy for continual learning. Preprint at *arXiv* https://arxiv.org/abs/1809.10635 (2019).
126. Kumaran, D., Hassabis, D. & McClelland, J. L. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends Cogn. Sci.* **20**, 512–534 (2016).
127. Qian, T. & Aslin, R. N. Learning bundles of stimuli renders stimulus order as a cue, not a confound. *Proc. Natl Acad. Sci. USA* **111**, 14400–14405 (2014).
128. Collins, A. G. E. & Frank, M. J. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychol. Rev.* **120**, 190–229 (2013).
129. Kirkpatrick, J. et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl Acad. Sci. USA* **114**, 3521–3526 (2017).
130. Zenke, F., Poole, B. & Ganguli, S. Continual learning through synaptic intelligence. *Proc. Mach. Learn. Res.* **70**, 3987–3995 (2017).
131. Masse, N. Y., Grant, G. D. & Freedman, D. J. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proc. Natl Acad. Sci. USA* **115**, E10467–E10475 (2018).
132. Zeng, G., Chen, Y., Cui, B. & Yu, S. Continual learning of context-dependent processing in neural networks. *Nat. Mach. Intell.* **1**, 364–372 (2019).
133. Bouchacourt, F., Palminteri, S., Koechlin, E. & Ostojic, S. Temporal chunking as a mechanism for unsupervised learning of task-sets. *eLife* **9**, e50469 (2020).
134. Harvey, C. D., Coen, P. & Tank, D. W. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature* **484**, 62–68 (2012).
135. Rule, M. E., O'Leary, T. & Harvey, C. D. Causes and consequences of representational drift. *Curr. Opin. Neurobiol.* **58**, 141–147 (2019).
136. Musslick, S. et al. Multitasking capability versus learning efficiency in neural network architectures. in *Annual Meeting of the Cognitive Science Society* 829–834 (Cognitive Science Society, 2017).
137. Bahri, Y. et al. Statistical mechanics of deep learning. *Annu. Rev. Condens. Matter Phys.* **11**, 501–528 (2020).
138. 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).
139. 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://arxiv.org/abs/1312.6120 (2014).
140. Seung, H. S., Sompolinsky, H. & Tishby, N. Statistical mechanics of learning from examples. *Phys. Rev. A* **45**, 6056–6091 (1992).
141. Goldt, S., Advani, M. S., Saxe, A. M., Krzakala, F. & Zdeborová, L. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *NeurIPS Proc.* (2019).
142. Jacot, A., Gabriel, F. & Hongler, C. Neural tangent kernel: convergence and generalization in neural networks. *NeurIPS Proc.* (2018).
143. Lee, J. et al. Wide neural networks of any depth evolve as linear models under gradient descent. *NeurIPS Proc.* (2019).
144. Advani, M. S. & Saxe, A. M. High-dimensional dynamics of generalization error in neural networks. Preprint at *arXiv* https://arxiv.org/abs/1710.03667 (2017).
145. Krogh, A. & Hertz, J. A. Generalization in a linear perceptron in the presence of noise. *J. Phys. Math. Gen.* **25**, 1135–1147 (1992).
146. Dauphin, Y. et al. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *NeurIPS Proc.* (2014).
147. Carey, S. Prēcis of 'The Origin of Concepts'. *Behav. Brain Sci.* **34**, 113–124 (2011).
148. Rogers, T. T. & McClelland, J. L. *Semantic Cognition: A Parallel Distributed Processing Approach* (MIT Press, 2004).
149. Walker, E. Y. et al. Inception loops discover what excites neurons most using deep predictive models. *Nat. Neurosci.* **22**, 2060–2065 (2019).
150. Schoups, A., Vogels, R., Qian, N. & Orban, G. Practising orientation identification improves orientation coding in V1 neurons. *Nature* **412**, 549–553 (2001).
151. Belkin, M., Hsu, D., Ma, S. & Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl Acad. Sci. USA* **116**, 15849–15854 (2019).

Acknowledgements

This work was supported by generous funding from the European Research Council (ERC Consolidator award to C.S. and Special Grant Agreement 3 of the Human Brain Project) and a Sir Henry Dale Fellowship to A.S. from the Wellcome Trust and Royal Society (grant number 216386/Z/19/Z). A.S. is a CIFAR Azrieli Global Scholar in the Learning in Machines & Brains programme.

Author contributions

The authors contributed equally to all aspects of the article.

Competing interests

The authors declare no competing interests.

Peer review information

Nature Reviews Neuroscience thanks M. Jazayeri 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.