Two views on the cognitive brain

David L. Barack and John W. Krakauer

Abstract | Cognition can be defined as computation over meaningful representations in the brain to produce adaptive behaviour. There are two views on the relationship between cognition and the brain that are largely implicit in the literature. The Sherringtonian view seeks to explain cognition as the result of operations on signals performed at nodes in a network and passed between them that are implemented by specific neurons and their connections in circuits in the brain. The contrasting Hopfieldian view explains cognition as the result of transformations between or movement within representational spaces that are implemented by neural populations. Thus, the Hopfieldian view relegates details regarding the identity of and connections between specific neurons to the status of secondary explainers. Only the Hopfieldian approach has the representational and computational resources needed to develop novel neurofunctional objects that can serve as primary explainers of cognition.

What is thought? Decision-making, planning, belief, recall, reasoning — all of these mental phenomena are about something. This fundamental and seemingly obvious insight has profound ramifications for the current state and future path of cognitive neuroscience. Contemporary cognitive neuroscience, especially animal model studies, often takes explanations of sensorimotor phenomena such as reflexes as the model. This strategy ignores the full implications of representational components to cognition. Instead, the best model of explanation for these cognitive phenomena relies on computation, the transformation of representations in the brain that result in behaviour. To go from movements to the mind, the explanation of intelligent behaviour requires a stronger notion of representation than the weaker one widespread in contemporary neuroscience used to explain sensorimotor phenomena.

The Sherringtonian view in contemporary neuroscience maintains that descriptions of networks of nodes, either neurons or areas of the brain and often including biophysical details about the neurons themselves, with specific weighted connections between them are needed to explain cognitive phenomena. Although this focus on molecules, cells and circuits may work for simple sensorimotor behaviours, we will argue that it fails to accommodate the semantic representations needed to explain cognition. In contrast to the Sherringtonian view, the Hopfieldian view emphasizes the role of neural spaces that explain behaviour in terms of computation and representation. Although these entities may result from the activity of neurons, ion flows and biomolecular processes, the Hopfieldian does not include details about them in explanations of cognition. Thus, unlike the Sherringtonian view, the Hopfieldian view starts at the level of representations and computations.

In the following, we will describe both of these views at greater length. Many researchers still adopt a Sherringtonian view, seeking to trace specific pathways in the brain and cataloguing types of cell. But advances are rapidly being made in understanding how representations can be realized by various forms of neural organization, especially populations. An emerging population doctrine provides support for Hopfieldianism and challenges the dominant neuron doctrine that inspires Sherringtonianism. Revolution is afoot, and an exploration is needed that outlines the structure of this emerging view. In this Perspective, we argue that the Sherringtonian view has limited explanatory resources for the description of the neural phenomena for representation and computation for cognition, whereas the Hopfieldian view is poised to reveal novel neural entities that restore representation as explanatorily dominant in explanations of cognition.

Neural explanations of cognition

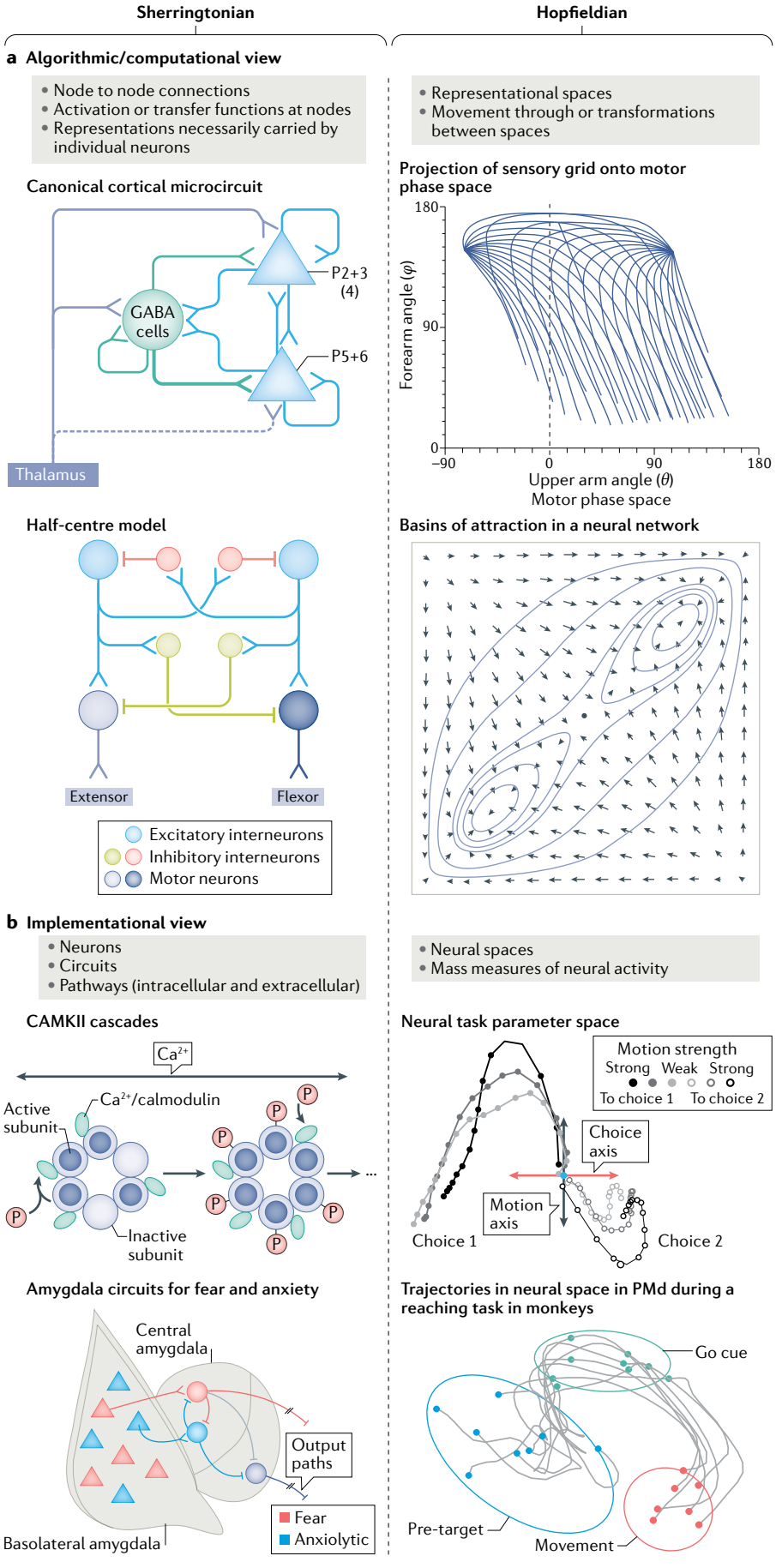
Cognition is computation over representations to yield behaviour[1]. Representation is a term widely used in neuroscience and refers to any informative, guiding neural signal. The signal carries information about states of the body or the external world. The information carried by these signals is used to guide behaviour.

In contrast to this neuroscientific usage, here we argue that cognition requires a more elaborate and restrictive notion of representation. Representations have content — they are about something. They are evaluable, such as for truth, success, accuracy and the like. They are detachable, capable of existing in the absence of their typical causes. They can be combined and interact in various systematic ways. Finally, they are produced and used by the system in order to generate behaviour. To be clear, these constraints on representation imply that not every sensory or motor state is a representation. For example, if a particular sensorimotor state cannot be activated in the absence of its typical cause, then that state is not a representation. In our view, cognition is the result of a restricted class of transformations of signals in the brain, ranging over only those signals that carry representational content in this more robust sense. The functions that underlie representational phenomena are fundamentally different in kind than simpler sensorimotor transformations and we argue that such functions require a fundamentally different ontology. It is at this juncture that philosophy becomes relevant.

We acknowledge this departure from the use of the term in contemporary neuroscience[2–10]. But how can such a constrained sense of representation be relevant to neuroscience? These properties should not be thought of as legislating the use of the term 'representation' but, rather, as a proposal about how to understand the computational role of the states posited by explanations of cognition in neuroscience. For example, the detachability of representations is a claim about the

causal structure of the nervous system. In particular, an explanation that requires that representations are detachable must be organized such that the presence of the representation is not stimulus bound: the representation can occur even in the absence of the stimulus in the world that is the representation's content. But this implies that any models that require stimulus input to cause that representation are wrong. Take, for example, the supposition — for which there is increasing evidence — that the inferotemporal cortex contains representations of objects[11–15]. These findings typically rely on the presentation of images (such as an airplane in flight) that initiates a processing cascade, which results in neural activity that correlates with the category of the image (such as an airplane). But a true representation of an object would be one that can occur in the absence of airplanes. We do not mean to imply that the inferotemporal cortex does not contain representations in this richer sense. Rather, the sorts of evidence that would establish this — such as the presence of these representations when planning a trip, imagining the plane flight or telling a story about last summer's vacation to Hawaii — remain to be gathered. Cleverer behavioural paradigms and computational models are clearly needed that can describe this behaviour. Further, more precise specifications are required for the other properties as well. Representations that match this more restricted understanding are semantic representations and the information that they carry are semantic contents.

We are concerned not just with representations but also computation. Computation includes how representations are transformed, updated, created or deleted. These transformations are often accomplished by information processing operations such as buffering, filtering and so forth. Such neural computations are performed over representations in the expanded sense and underlie cognition, resulting in changes in overt behaviour or internal systems. The transformations of representations are essential for understanding cognition[16] but are often overlooked (see, for example, REFS[17,18]). But such transformations cannot be ignored. To explain how behaviour is generated, it is not enough to explain the representations; the transformations must be identified and their neural realizations described. Cognitive neuroscience is focused on explaining cognitive phenomena conceptualized in terms of computation over representations in the brain.



◀ Fig. 1 | **Comparing and contrasting the commitments of the Sherringtonian and Hopfieldian views.** Description of the two views outlined, the Sherringtonian view and the Hopfieldian view. The Sherringtonian view is committed to explaining cognition as the transformation of signals by nodes in a point to point architecture. These nodes and connections correspond to neurons embedded in circuits and pathways in the brain. At the algorithmic/computational level, the Sherringtonian view (part **a**, left) explains cognition as the result of specific patterns of node to node connections where individual nodes transform representations. The canonical cortical microcircuit (top) is an example of such a specific stereotyped pattern, where input received by pyramidal cells (P) in layer 4 (L4) neurons is transformed and projected to L2/3 and then passed to L5/6. A second example is the half-centre model of the reflex (bottom). Excitatory interneurons and motor neurons exhibit stereotyped connections to interneurons for reciprocal and cross-inhibition. At the implementational level, the Sherringtonian view (part **b**, left) details neurons, circuits and intracellular and extracellular pathways that implement the circuits at the algorithmic/computational level. Intracellular CAMKII cascades, for example, are persistently active molecular cascades within cells due to autophosphorylation. Specific circuits for fear or anxiety in the amygdala involve different patterns of neuronal connectivity between the basolateral and central amygdala, with dedicated output paths for each affective process. The Hopfieldian view, in contrast, is committed to representational spaces with computation cast as the transformation between or movement within those spaces. Networks of neurons and mass measures of neural activity implement those spaces and transformations. At the algorithmic/computational level, the Hopfieldian view (part **a**, right) describes representational spaces and the way that cognitive systems move through them or transform them. A simplified toy example describes how projection from a space encoding eye position in retinal coordinates metrically deforms that space to encode arm movements in upper arm and forearm angle space (top; REF.[129]). Basins of attraction in neural networks that serve as a substrate for memory or decision-making also illustrate the view (bottom; REF.[60]). At the implementational level (part **b**, right), these representational spaces and transformations are implemented in neural spaces assessed using mass measures of neural activity such as population recordings of many neurons. In a perceptual decision-making task, neural trajectories emerge when neural activity in the prefrontal cortex is plotted in a task-variable space composed of a choice axis, representing the direction chosen, and a motion axis, representing the degree of coherence (strength of evidence) of a field of randomly moving dots (top; REF.[75]). Neural activity projects further along the motion axis to represent the strength of evidence in the dot motion stimulus. Neural trajectories are also observed in dorsal premotor (PMd) neurons for motor planning (bottom), where the population activity passes through the same sequence of states in each trial (light-grey lines) in order to set movement parameters in a downstream population. Part **a** (top left) adapted with permission from REF.[128], Elsevier. Part **a** (bottom left) adapted with permission from REF.[51], Elsevier. Part **a** (top right) adapted from REF.[129], Springer Nature Limited. Part **a** (bottom right) adapted with permission from REF.[61], PNAS. Part **b** (top left) adapted from REF.[130], Springer Nature Limited. Part **b** (top right) adapted from REF.[75], Springer Nature Limited. Part **b** (bottom left) adapted from REF.[131], Springer Nature Limited. Part **b** (bottom right) adapted from REF.[132], Springer Nature Limited.

The Sherringtonian view

The Sherringtonian view defined

The Sherringtonian view of cognitive explanation emphasizes the specific connections between neurons in the brain (FIG. 1). The Sherringtonian view maintains that cognition will be explained just as Sherrington explained reflexes[19,20] (see also REF.[21]). For the Sherringtonian view, neuron to neuron connections and the computations performed by these neurons or the traceable circuit in which they are embedded are the first-level explainers of cognition[22] (BOX 1). The Sherringtonian view has been essential to twentieth-century neuroscience, providing insights into a range of CNS phenomena.

At the implementational level, the Sherringtonian view describes, in biophysical and physiological terms, the neurons and connections that realize a cognitive phenomenon. These descriptions include specific neural transfer functions: the transformations performed by single neurons over their inputs, typically in the dendritic tree or the axonal hillock. They also include information about particular neurotransmitters, such as their differential role in neural signalling, and the details of local circuit connections: which neurons are connected to which others. The first-level explainers of cognitive phenomena are circuits made up of particular neuron to neuron connections realized by specific neurons with fixed biophysical identities and utilizing particular neurotransmitters to pass signals between them.

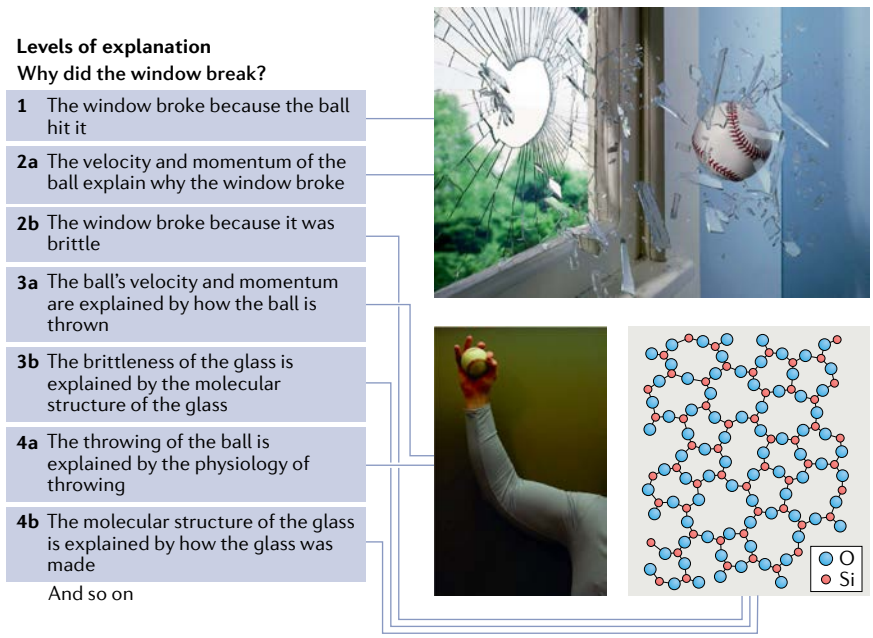
At the algorithmic level, the Sherringtonian view appeals to computations performed by networks of nodes with weighted connections between them. In the brain, these nodes are neurons and these connections are synapses. The neurons perform dedicated computational transformations over signals received from other neurons in the network. Explanations of cognition are described in terms of information intake by individual cells, transformation of that information via neural transfer functions and then output to other cells. This computational point goes beyond mere implementation because there may be different implementations of the same circuit[23]. The Sherringtonian view can omit biophysiological details such as types of cell, biomolecule, neurotransmitter and so on in the algorithmic description of the computations and representations, and so their descriptions can satisfy 'medium independence' (see BOX 2). This framework assumes that cognition is as amenable to a Sherringtonian form of algorithmic abstraction as are reflexes (reciprocal inhibition), eye movements (the neural integrator) and central pattern generators (the half-centre model). Perhaps the most notable example of the algorithmic Sherringtonian view is the attempt to explain cortical activity in terms of a single canonical circuit: a specific pattern of connectivity between neurons in different cortical layers that was first proposed in the 1970s to explain cortical activity in all its diversity[24–26]. Thus, in the Sherringtonian view, explanations of cognition will always take the form of computations performed by individual neurons and signals passed over their connections.

Illustration of the Sherringtonian view

The type of explanation present in studies on motion perception and discrimination illustrates the Sherringtonian view[27–30]. This research programme investigates how monkeys transform perceptual signals from a pattern of coherently moving dots embedded in random dot noise into a decision about the dominant direction of motion of those dots[31]. The capacity to discriminate motion direction is decomposed into a series of processing steps with signals transformed by single neurons and then passed on to the next neuron in the processing chain[32,33] (FIG. 2). Area V5/MT contains neurons that are sensitive to the speed and orientation of motion in a visual stimulus[34–36]. Consider a single neuron in this area. In the Sherringtonian view, this neuron is part of a local circuit for computing the speed and direction of motion stimuli. The neuron contributes to the circuit by being connected to specific upstream and downstream neurons, and by performing a transformation over the incoming signals it receives to then pass that transformed signal along to the next processing step. These transformations are implemented by particular types of synaptic operation (such as integration) and transmitted by specific signalling molecules (presumably glutamate in the case of large pyramidal cells) between cells and

Box 1 | **Levels of explanation**

The two views on how to explain the neural basis of cognition fundamentally disagree over the content of first-level explanations of cognition (see the figure). A level of explanation is the relative importance placed on entities in an explanation. In explaining why the window broke, the speed and direction and hardness of the ball are more important than the microphysical constituents of either the ball or the window. If one wants, instead, to explain the speed and hardness of the ball, these two properties will be explained in terms of how the ball was thrown, say, and how the molecules of the ball are organized, respectively, thereby placing those details at the next explanatory level; and so on for how the ball was thrown and why the molecules are organized as they are, whose explainers will be at the third level. For a given phenomenon, the first explanatory level refers to the entities and properties that are referred to in explanation; the second level refers to the entities and properties that explain the first-level entities and properties; and so on. The ability of second-level explainers to explain first-level explainers will not confer upon them first-level explainer status. Each level explainer is essential and not just a placeholder. However, this is not to deny the importance of second-level explainers in explanations of first-level phenomena. In this view of explanation, the two views posit different first-level explainers for cognition. The Sherringtonian view maintains that the first-level explainers will be individual neurons and their detailed connections. The Hopfieldian view maintains that the first-level explainers will be properties of neural spaces.



intracellular ion flows (such as Ca^{2+}, K^{+}, Cl^{-} and so on) or large biomolecules within cells.

As part of this processing stream, MT neurons project to neurons in the lateral intraparietal area[37]. As motion evidence is displayed to the subject, neurons in the lateral intraparietal area show a stereotyped pattern of rising activity that crescendos just prior to the initiation of an eye movement to indicate the direction of motion[29]. This activity has been modelled as the integration of evidence towards a bound to make a decision about the direction of motion[33]. Although these findings have recently been challenged (see, for example, REFS[38,39]), this research programme still nicely illustrates the Sherringtonian model for explaining cognition.

In sum, the animal's ability to make perceptual decisions in noisy sensory conditions is explained by neurons summing up motion evidence by integrating perceptual signals received from neurons in motion processing regions. Generalizing, a circuit — conceptualized as a series of processing steps each performed by individual neurons that are appropriately connected together — executes the needed computations. One could imagine a similar circuit for a more abstract operation, such as where to gather information, how to make long-term plans or how to select between two further decisions. The Sherringtonian view is committed to the explanation of cognition via point to point communication between neurons organized into circuits.

Problems with the Sherringtonian view

Recently, the Sherringtonian view has come under heavy criticism. One attack on Sherringtonian-style cognitive neuroscience charges that the approach cannot explain combinations of stimulus properties[40]. This charge originates in a critique of the first generation of artificial neural networks called perceptrons[41] that demonstrated that these simple artificial neurons cannot perform the logical operation of exclusive or (XOR)[42]. For example, suppose we need to categorize stimuli as either cats or dogs but not both. Linear feedforward one-layer artificial neural networks cannot implement a function for this simple exclusive disjunction categorization problem. The problem of exclusive disjunction has become one of the most important motivations for the Hopfieldian view[40,43] (FIG. 3c).

As a critique of perceptrons, the XOR challenge is on sure ground. The issue is not with perceptrons, however, but with biological neurons. Note, however, that there are no conceptual a priori arguments against single neurons executing XOR computations. For example, suppose neurons A and B both synapse on neuron C. Suppose the weight on A is –1 and the weight on B is +1. Suppose neuron C fires if and only if the absolute value of the activity is greater than +0.5. Suppose inputs are additive, and active neurons are in state 1, inactive neurons in state 0. Then, when both neurons A and B are active, their inputs will cancel; but if only one neuron is active, then neuron C will be active. As a conceptual point then, there is no bar to XOR computations over inputs to single neurons.

The explanation by appeal to specially connected triplets of cells does not scale: for every possible XOR categorization, there will need to be a dedicated circuit (that is, a giant look-up table), which is wildly impractical[1]. Furthermore, because of the absolute value operation, computing XOR is only possible with a non-linear function, although nothing about the Sherringtonian view implies that single cells can perform only linear operations. Nonetheless, the example does suggest something interesting: for some subset of XOR problems, assuming non-linear read-outs and positive and negative weights, there could be dedicated neurons. The conclusion stands even if we relax those assumptions, although other assumptions would be needed. This point about the complexity of computation in single neurons has not escaped the literature[44,45], and, at least in ex vivo preparations, XOR computations have been observed in hippocampal neurons[46,47] and in human neurons from cortical layers 2/3 (REF.[48]).

A host of population-level phenomena are often cited in challenge

to the importance of neuron to neuron connections[17]. These include reverberating activity[49], cell assemblies[50], central pattern generators[51] and even oscillatory communication between brain areas[52]. In addition, many novel models of neural data involve uncovering low-dimensional latent structure from high-dimensional data[53–55]. However, the Sherringtonian approach can sanction population-level phenomena. Various forms of population codes or computations are acceptable as long as each neuron is committed to a particular role in the circuit on account of its properties. The importance of embedded latent low-dimensional dynamics can also be made consistent with the Sherringtonian view if the specific neuronal circuitry that gives rise to those low-dimensional dynamics is taken as the first-level explainer. For example, an active research programme investigates whether the dimensionality of neural populations can be estimated on the basis of the distribution of local circuit motifs[56–59]. In the Sherringtonian view, groups of neurons can play a role in computation, but they will not be the first level in the explanation of cognitive phenomena. The key explanatory work is performed by single neurons and their interconnections. Population-level computational phenomena are a strike against Sherrington only if attempts to account for them in local circuit terms fail.

The Hopfieldian view

The Hopfieldian view defined

In contrast to the Sherringtonian view, the Hopfieldian view emphasizes the distributed nature of computation for cognition in neural systems just as Hopfield illustrated how distributed neural networks could perform computations[60–62] (see also REFS[50,63–65]). The approach couches its operations and representations in terms of transformations between neural spaces. Implementationally, massed activity of neurons is described by a neural space that has a low-dimensional representational manifold embedded within it[66]. These neural spaces may be comprised of neural ensembles[50,67,68], brain regions[67,69] or distributed representations across the brain[70]. These representations and transformations are realized by the aggregate action of neurons or their subcomponents, but explanations of cognition do not need to include a biophysiological description of neurons or their detailed interconnections. Single neurons can play a role only as second-level explainers of cognitive phenomena, explanatory only by virtue of their contributions to neural spaces. In an extreme form, the Hopfieldian view avoids single cell details altogether (see, for example, REF.[71]).

Algorithmically, Hopfieldian computation consists of representational spaces as the basic entity and movement within these spaces or transformations from one space to another as the basic operations. The representations are basins of attraction in a state space implemented by neural entities (be they single neurons, neural populations or other neurophysiological entities) but agnostic to implementational details (although, as a matter of fact, most Hopfieldian computations are focused on neural populations). A space of parameters describes the dimensions of variation of the representational space. This view of representation shares a conceptualization of content with quality-space approaches in philosophy[72,73]. The computations over those representations are transformations between spaces or movement within them and are described in terms of the dynamical features of representational spaces such as attractors, bifurcations, limit cycles, trajectories and so on. In short, cognitive functions are realized by neural spaces and the system's movement within or between them[16,18,40,66].

The Hopfieldian view illustrated

In the Hopfieldian view, the explanatory role of a single MT neuron is the result of its membership in a population that implements the neural state space used to represent motion. Visual input activates or suppresses this neuron such that its activity helps shape the way that the neural space changes over time. This multidimensional neural activity space is the result of the firing of individual cells and other non-spiking features of their neural activity such as membrane dynamics and neuronal correlations. A representational manifold is embedded in this neural space. The point inhabited by the system in this space is the representation of the speed and direction of motion. The computations underlying cognition are also organized only at the level of the neural space. For example, consider again the lateral intraparietal area's role in noisy perceptual decision-making. This region, in fact, exhibits a diversity of responses that betray the Sherringtonian story above and may suggest a population-level computation of integration and representation of evidence[74]. Many neurons do not exhibit integration, although if neural activity from those neurons is combined, the pattern of integration is evident. On a similar task, evidence integration was present in a low-dimensional projection of neural activity recorded from the prefrontal cortex into a space defined by task-relevant variables — that is, a representational space[75]. In the Hopfieldian view, the neurobiological and the representational are integrated because lower-dimensional representational spaces are embedded in higher-dimensional neural ones.

The emphasis in the Hopfieldian view on representational spaces provides explanatory resources unavailable to the Sherringtonian view. Consider the explanation of errors in a simple memory task[76,77] (FIG. 3). In this task, subjects first observe a stimulus or a sequence of stimuli, followed by a delay and then presentation of further stimuli. The subject must remember whether these later stimuli match the initially presented ones and then select the matching stimulus

Box 2 | **Marr's three levels of analysis**

Marr[133] described three levels at which a system can be analysed. Here, we present a modified take for cognition in this classic view. At the first level, which Marr denoted the 'computational' level but which we call the 'ecological' level[134], the problem facing the organism and how the organism solves that problem is described. What are the variables the organism must track and how must they be transformed? Why track those variables and transform them in that way? This is a description of what the organism must do and why it must do it in order to solve this problem[135]. At the second level, the algorithmic level, the procedure that carries out the computation specified at the ecological level is described. What are the functional properties of the organism and how do they track and transform the variables specified at the ecological level? What are the basic operations in the system (such as if–then statements, for loops and so on)? How are the basic operations combined to execute more complex operations? In what order are the basic operations or the composed functions to be performed? The algorithmic level specifies the basic set of representations and computations used to process information in the system and how those operations are combined (the architecture). Importantly, the architectures specified at the algorithmic level can be realized by different types of physical system such as computers or brains (medium independence[136]). At the third level, the implementational level, the physical realization of the algorithm is described. What is the system made of? What are the physical parts of the system, what do those parts do and how are the parts and activities organized? The implementational level describes the physical processes whose activity realizes the operations described at the algorithmic level.

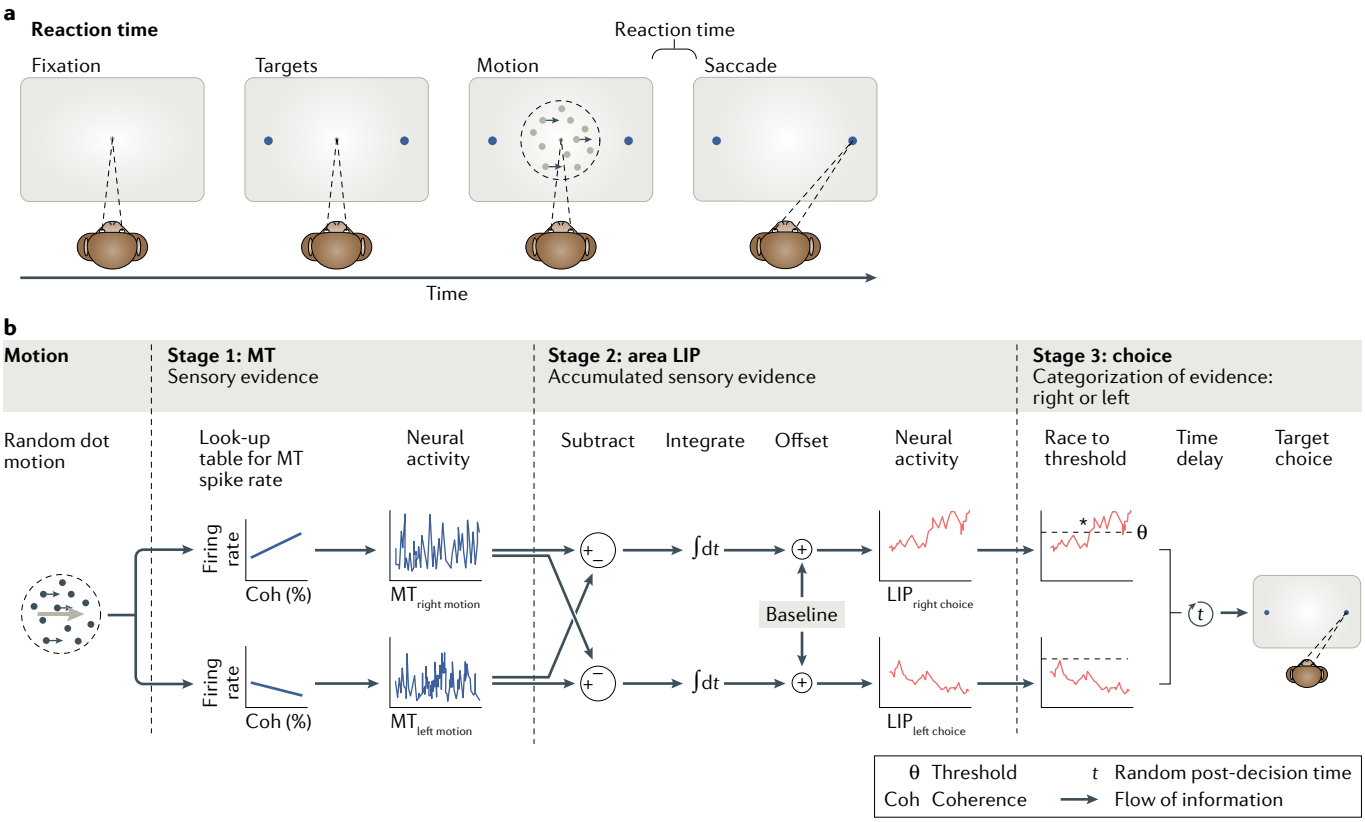


Fig. 2 | **Depiction of how the motion direction decision-making research programme illustrates the Sherringtonian view. a** | Random dot motion task. Monkeys first fixate and then maintain their gaze as two targets appear, followed by a centrally presented field of randomly moving dots, some fraction of which move coherently. Monkeys then indicate their decision about the direction of motion with an eye movement to a target in the corresponding direction. **b** | Explanation of the decision in Sherringtonian terms. A representation of the motion evidence, the evidence about the direction in which the dots are moving, is carried by neurons in area MT. This signal is then sent along dedicated pathways to the lateral intraparietal area (area LIP). Neurons in area LIP integrate the motion evidence from MT to form a representation of the sum of the evidence. This representation thresholds, signalling downstream action selection and initiation. Motion: random dot motion that varies in direction and strength (% of dots moving in same direction) is displayed to the monkey. Stage 1, MT: modelled neurons in area MT exhibit a noisy firing rate that correlates with the direction and strength of motion. Stage 2, area LIP: modelled neurons in area LIP integrate motion signals received from the modelled neurons in MT. Area LIP neuronal input is the summed input of the difference between left-preferring and right-preferring MT neurons. Area LIP neurons show response preferences that match the MT neuron direction of motion preferences. This input is summed over time, a baseline offset is added and, then, the modelled neurons exhibit a noisy signal that correlates with this sum. Stage 3, choice: left-choice and right-choice area LIP neuron firing rates rise in proportion to their input, with the first pass a threshold determining the choice of target. Parts **a** and **b** adapted with permission from REF.[32], Oxford University Press.

from an array of non-matching distractors. Neural recordings from the lateral prefrontal cortex performed during this task were first analysed by decoding either the type of task (recall or recognition) or the cues used in the task[78]. In a decoding analysis, a network is trained to classify some stimulus or condition on the basis of some neural measure such as the observed activity of neurons. The analysis used included only purely selective neurons, those that showed a main effect in their firing rate of only one task variable such as a cue or condition. Next, another decoding analysis was performed on the portion of the population that showed only mixed selectivity, neurons whose firing rates were modulated by combinations of conditions or cues. In both cases, decoding accuracy was above chance. The dimensionality of the population activity, the number of independent ways that the population activity can vary, was then estimated. This dimensionality was estimated to be higher when the mixed selectivity was included. Importantly, the neural population showed lower dimensionality on error trials than on correct trials, a difference that disappeared if non-linear mixed selectivity was left out of the decoding. The collapse of the dimensionality of the population on error trials reflects a reduced representational capacity in the population and helps explain the errors committed on those trials. In sum, the number of independent dimensions needed to describe population activity is used to explain the cognitive behavioural phenomenon. Non-linear mixed-selectivity neurons are needed because they give rise to more complex representational spaces. This explanation is an example par excellence of the Hopfieldian approach.

Problems with the Hopfieldian view

The Hopfieldian view faces a number of its own apparent difficulties. Examples abound of information carried by single neurons, including sensation (for example, tonotopy in A1, interaural timing difference computations in audition, somatosensory specificity or numerous findings throughout the visual system), navigation (grid cells in the entorhinal–hippocampal cortices) or learning (such as temporal difference prediction errors conveyed by dopamine cells). Consider as an example the role of grid cells in the entorhinal and hippocampal cortices in navigation[79,80]. Grid cells are neurons that tile the space within which the animal finds itself (see, for example,

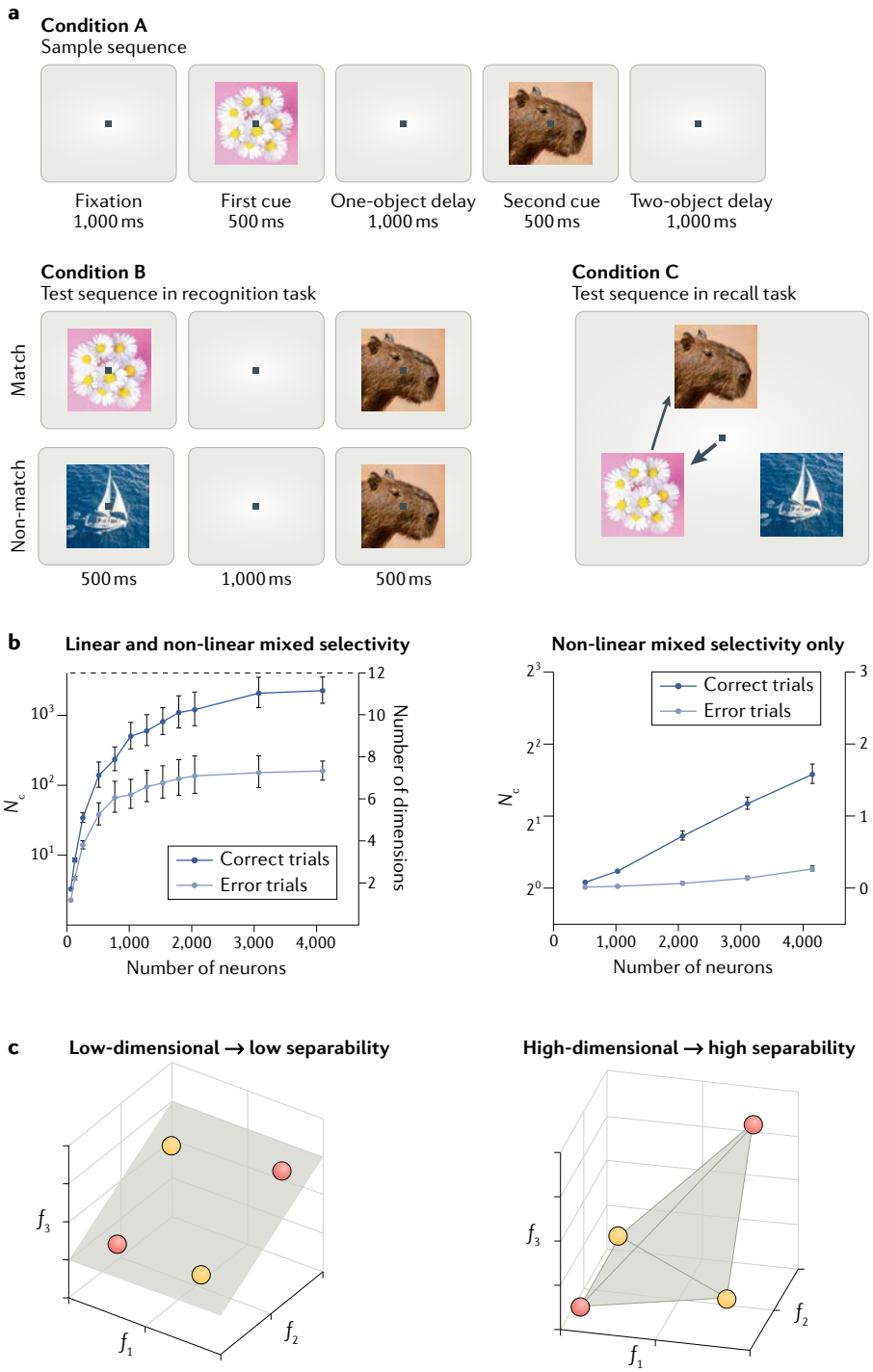


Fig. 3 | **Depiction of how the delayed-recall research programme illustrates the Hopfieldian view. a** | Delayed match to sample task. Condition A: sample trial sequence for the task. Monkeys first fixate on a central square for 1,000 ms. A sample image is then shown for 500 ms, followed by a 1,000-ms delay. A second image is shown for 500 ms, followed by a second 1,000-ms delay. Monkeys are then required to respond in different ways depending on the task. Condition B: recognition task sequence. After the second delay in condition A, monkeys are shown a second sequence of two images. The first image appears for 500 ms, followed by a 1,000-ms delay then the second image for 500 ms. The match condition occurs when the second sequence of images matches the original sequence in Condition A; the non-match condition occurs when the second sequence fails to match the original. Monkeys must indicate whether the second sequence is a match or a non-match. Condition C: recall task sequence. After the second delay in condition A, three images are shown. Monkeys must select the two images displayed in the original sequence in condition A in the order in which they were shown. **b** | Dimensionality (that is, number of dimensions) of the representational space is correlated with the number of neurons and whether the trial was correct or an error. This dimensionality is a feature of the population and not reducible to individual neurons; hence, this explanation is not available to the Sherringtonian view. Left panel: correct trials have a larger dimensionality than incorrect trials across a range of population sizes. Number of neurons used in the analysis is on the *x* axis, and number of binary classifications (N_c) is on the left *y* axis. The number of binary classifications is the number of task conditions that a classifier can be trained to classify (1 = trial was condition C; 0 = trial was not condition C) for a given performance threshold of correct classifications (in REF.[78], this threshold varied in the range 75–80%). The number of dimensions on the right *y* axis is a logarithmic function of the number of binary classifications. Right panel: same as left panel but the linear contribution of each neuron has been removed. Correct trials still display an increase in dimensionality compared with error trials. **c** | Increases in dimensionality are required for basic logical operations (f1, f2, f3 are dimensions of the representational space). Left panel: a plane cannot be used to classify items using an exclusive or rule — there is no line that can be drawn on the plane that separates the yellow items from the red. Right panel: shifting to three dimensions allows for a linear classifier to correctly partition the items. Parts **a** and **b** adapted from REF.[78], Springer Nature Limited. Part **c** adapted with permission from REF.[40], Elsevier.

REFS[81–83]). These cells (and many others in the medial temporal lobe) seem to have particular representational contents that are proposed to play a central role in explanations of spatial cognition and navigation, constituting a cognitive map of the environment[84]. They are also proposed to play a role in internally directed cognitive search[85,86]. Such single-neuron representations appear to be at odds with the focus on populations at the heart of the Hopfieldian approach.

The Hopfieldian view, however, denies the explanatory power of approaches that ascribe computations rather than just informational correlations to single cells. The selective responses of single neurons are explanatorily derivative. Instead, neural spaces play the central representational role. Note, however, that the appeal to the presence of population activity is insufficient to defend against the objection posed by place cells, grid cells and other selective responses in single neurons.

As pointed out above, the Sherringtonian view can welcome population codes because, in that account, these codes are explanatory of navigational abilities by virtue of the single-neuron constituents and interconnections that comprise the

population. The Hopfieldian view has to maintain that the order of explanation is the reverse of the Sherringtonian view: the single-neuron responses possess explanatory power only because they are part of a larger population. At best, single-neuron findings have an explanatory role only at the second level, as explanations of those neural spaces. Furthermore, single-neuron correlations do not imply that the relevant computation is occurring at that level. Correlation does not imply computation. For example, the activity of single neurons, even activity that correlates with behaviourally relevant internal or environmental variables, does not imply a computational role for that activity because such correlations can have other non-cognitive biological functions. Also, noisy complex systems such as the brain often contain spurious correlations. Nonetheless, some single-cell responses can provide a clue to what is being represented.

The Hopfieldian view can, for example, accommodate grid-cell responses by arguing that they are the result of population-level processing that sculpts their activity by training and feedback. In this view, a learning process creates a neural space that naturally results in the formation of grid cells. This neural space is the result of the solution to the ecological problem facing the organism, which is the problem of coding the position of the agent. For example, the optimal solution to path integration corresponds to a ring in Fourier space[87]. Once this ring is implemented, grid cells develop in the network[88,89].

Both replies illustrate a more general tactic that the Hopfieldian view can rely on. Cognitive state spaces are embedded in high-dimensional neural spaces. Activity in any given cognitive space could predominantly drive single neurons. On this tactic, significant single-neuron activity may in fact emerge from the population activity. The Hopfieldian focus on the neural and representational state space contends that populations perform computations and constitute representations by virtue of general principles of operation of neural networks and not by virtue of detailed connectivity profiles of the neurons that make up the populations. As a result, the Hopfieldian view stresses aggregate neural activity and contends that the explanatory power of single neurons derives from their membership in the population and not from the connections to or from that neuron. The Sherringtonian view, by contrast, must maintain that population activity is constituted and driven by the activity of single cells and their specific connections. For the Sherringtonian view, the connections matter; for the Hopfieldian view, they do not. Another way to intuit the difference between the approaches is to realize that trajectories through state spaces are invariant to the specific neurons sampled from the population to generate them. Cognition lies at a level above a one-to-one correspondence between the neural space and a particular pattern of connections between a specific set of identified neurons.

In addition to single-neuron phenomena, numerous studies have revealed specific patterns of connectivity between regions. Wiring diagrams for the brain are fantastically complex, featuring dozens of regions with specific pathways between them[90]. Indeed, the determination of which areas are wired to which others may be genetically determined, and this may play an important role in the explanation of various capacities such as those for vision[91]. The Hopfieldian view must also account for this specificity. If population dynamics are the first-level explainers, then why does the brain exhibit such specialized wiring?

The Hopfieldian view has a few replies at the ready. The best bet for the Hopfieldian view may be to respond that the wiring diagram is a second-level explainer of visual processing. The first-level explainer is the population dynamics of each region, and the wiring helps explain those dynamics. Another way to accommodate specific connectivity between areas is to permit that populations need to be connected to each other but without the need to commit to full specification of the neuron to neuron connections within or between areas. The Hopfieldian view emphasizes activity in neural spaces but can remain agnostic on how information is transmitted between such spaces. There is no need to adhere to an extreme form of Hopfieldianism that rejects all regionally specific functions in the brain[70] (for an extended discussion, see REF.[92]).

Possible paths to unite the views

Cognitive neuroscientists are interested in characterizing the functional architecture of the brain[93,94]: the fundamental computational operations performed over representations that the brain carries out for cognition. The Sherringtonian and Hopfieldian approaches present different hypotheses about how best to describe and to uncover the brain's functional architecture. There are different ways to analyse the relationship between the two views.

Resolution of the debate

The whole debate between the Sherringtonian view and the Hopfieldian view discussed here could be construed as old news if framed in a particular way. In this framing, the Sherringtonian view's emphasis on neuron to neuron connections contrasts with the Hopfieldian view's emphasis on neural populations and relates to the debate between the neuron doctrine[20,50,63,95–99] and the population doctrine (as it is now called; see, for example, REFS[18,66]). The neuron doctrine was originally formulated in opposition to reticularism[63], but modern incarnations emphasize the role of the single neuron in information processing in the brain. The divide then boils down to a clash between the neuron doctrine that maintains that single neurons are the basic explanatory unit for cognition and the population doctrine that maintains that the central explanatory role will be played by neural populations.

We reject this rather pedestrian version of the debate. The Sherringtonian and Hopfieldian views, as formulated herein, are updated to provide a more plausible take on the role of single neurons or populations. In fact, both the Sherringtonian and Hopfieldian views acknowledge the importance of single neurons and neural populations. Take the Sherringtonian emphasis on single neurons first. The claim that the Sherringtonian view is false because of the need for more than one neuron is neither charitable nor accurate. No one ever believed that the visual scene can be captured by a single retinal cell. The most charitable interpretation of the neuron doctrine is not the idea that individual cells, and only individual cells, are the explainers of cognitive, or any behavioural, phenomena. Rather, it is the neuron+ doctrine that individual cells, and their interconnections, are the explainers of cognitive phenomena — this epitomizes the 'circuit-cracking' approach so prevalent in current neuroscience. The neuron+ doctrine can accommodate the findings regarding neural assemblies by stressing that the specific connections underlying those assemblies will inform their function. The neuron+ doctrine is just the Sherringtonian view where neuron to neuron circuit specifications are needed for explanations of cognition, just as has been achieved for simpler sensorimotor phenomena. Further, insofar as neural populations are central to understanding cognitive phenomena, advocates for Sherrington-style neuroscience can emphasize that neural populations play their roles by virtue of the connections

and transformations performed by their single-neuron constituents.

The Hopfieldian view, in turn, can acknowledge the importance of single neurons. Undeniably, single neurons contribute to population activity. However, the specific connections between cells are irrelevant as first-level explainers of cognition. The most charitable interpretation of the Hopfieldian view presents a population+ doctrine that populations of cells constructed from individual neurons are the explainers of cognitive phenomena. As a result, the population+ doctrine, which is just the Hopfieldian view, can assimilate the activity of keystone cells as the predominant drivers of population activity in particular contexts. Cognitive phenomena result from the action of populations, whether driven by single cells or as a result of the effect of many cells, and it is only when we turn to explain these population phenomena themselves that single neurons and their point to point organizations take prominence.

Thus, the characterization of Sherringtonianism and Hopfieldianism as merely a contrast between the neuron and population doctrines is incorrect. Neither view denies that both neural populations and single cells are important and this importance can be reconciled with either theoretical commitment. Consequently, the mere involvement of neurons or populations fails to determine a dominant view, and we find the contrast between single neurons and neural populations to be facile. Does the possibility of including single neurons or populations in either view suggest that the views are closer than we have illustrated? Can their disagreements be reconciled?

Replacement or reduction

In a replacement vein[66], the Hopfieldian approach presents a new population+ doctrine that will replace the older Sherringtonian approach. In Kuhn's classic model[100], revolutionary science occurs when a field is faced with problems intractable under current theories or doctrines. Both exclusive disjunction[40,42,43] and the diversity of prefrontal cortical responses[75] challenge Sherrington-style neuroscience. In response, a population+ doctrine, the Hopfieldian view, has been formulated with its own types of analyses, tools, techniques and concepts.

In reply to the possibility of revolution, however, the Sherringtonian view could instead attempt to reduce the phenomena underlying the Hopfieldian approach to cell to cell connections. This option would require the development of novel explanatory resources to resolve the challenges facing the Sherringtonian view, including resolution of the mixed selectivity and exclusive disjunction problems. But, fundamentally, the response to the Hopfieldian challenge is to hold out hope that for each way of performing a cognitive operation there will be a description in terms of neurons and local circuit connections. The effort to study the neural basis of cognition in insects with an impressive array of new tools can be seen as a move of this type[101,102]. However, such an approach faces the challenge of accounting for the semantic components of cognition by appeal to what are, essentially, second-level explainers. Needless to say, we are sceptical of the viability of this strategy.

Reconciliation

Instead of replacement or reduction, the differences in the approaches may reflect differences in brain areas due to evolution or to the types of function performed by different regions. The Sherringtonian approach may be more effective for older, conserved, or modular structures such as the brainstem or spinal cord. Sherringtonian circuits are computationally dedicated modules that reflect the outcome of selective evolution. Neuron to neuron connections and canalized local circuitry would reflect one outcome of such selection. Newer, flexible or recent structures might require the Hopfieldian approach. Hopfield circuits are more flexible modules that can be used for a range of computations. These functions do not result from evolutionary selection and so would be implemented by brain areas either without dedicated functions or with specializations that can be recycled for novel information processing for new challenges[68,103]. Distributed population activity for computation would reflect the outcome of neural processing whose goal is to approximate those new but necessary computations. The recalcitrance with which the cerebellum and basal ganglia, for example, have refused to reveal their universal function may be due to conflicting evidence resulting from applying both Sherringtonian and Hopfieldian approaches to the question.

Second, the differences in the approaches may instead conceal common ground that may bridge the two views to address the shortcomings without relinquishing the successes of both. Some cognitive neuroscientists describe a dictionary of functional types that might underlie neurocognitive computation. These include Fuster's cognits[104], Arbib's brain operating principles[105] or Carandini's canonical neural computations[106]. Several approaches in philosophy and neuroscience that emphasize dynamic coalitions of neurons are reminiscent of this approach as well[67,68]. A succinct example of this approach is embodied by Wang's research on reverberatory dynamics[49,107]. The synaptic reverberation model of perceptual decision-making, now extended to many different cognitive functions[108,109], describes one such basic functional type and is applicable to both individual neurons as well as neural populations. However, as the functional types must be combined into complexes to provide an explanation of the targeted cognitive phenomenon, at best they would be second-level explainers.

Revolution

A revolutionary cognitive neuroscience would go a long way to explain and dissolve the tensions between the Sherringtonian and Hopfieldian views. Both views focus on relevant neural phenomena for different cognitive tasks or for different brain regions. The way to unify the two views is to understand that there are underlying computational entities implemented by the brain. In some cases, those implementations take the form of single neurons and specific connectivity profiles. These circuits inspire the Sherringtonian approach. In other cases, those implementations take the form of mass activity in neuronal populations that disregards single neurons and circuits. These neural spaces inspire the Hopfieldian view. But both are instances of the implementation of a computation that is used to transform representations for behaviour.

The characterization of this novel cognitive neuroscience has its best springboard in Hopfieldianism. The Sherringtonian approach assumes a level of detail that restricts its available explanatory resources, whereas the Hopfieldian approach adopts fewer such constraints. These restrictions are manifest both representationally and computationally.

Computation and Sherringtonianism.

There is a deep representational critique of Sherringtonianism that the Hopfieldian view avoids. The Sherringtonian view needs to explain cognition as the result of dedicated neural circuits. Much like explanations of the reflex that appeal to a particular circuit performing its function, cognition results from dedicated circuits performing theirs. This implies that only the individual nodes can be assigned semantic contents. There are two problems with this.

Suppose each node is assigned individual contents. Because nodes carry the semantic content and there are different semantic contents that can cause any given movement, then for each such behaviour, the Sherringtonian view requires a dedicated circuit that must contain distinct nodes. For example, not all hand waves result from the same cause even if the movement is the same across all of the different instances; a wave for help is not the same as a wave hello, and in this approach each such wave requires a distinct circuit[94]. But for each representational way to cause a movement, a new neuron to neuron reflex-like pathway is required, with the end result of an explosion in circuit resources[1]. Alternatively, multiple contents could be assigned to the same node. Say the range from *m* to *n* of the firing rate of this neuron means 'bear', from *n* to *r* means 'briar' and so on. This increases the complexity of semantic organization of the system. The explanation of cognition requires the brain to transform representations in light of their contents, which means that representations must be used[9], as captured by our requirement above that representations are produced and consumed by the system. Such a complex semantic organization, however, may not be easily usable by the brain.

In addition, the representation of complex contents is problematic for the Sherringtonian view. Consider the assignment of the content that snow is white to a single node. This is not, in principle, false — early connectionist networks sometimes assigned explicit contents to nodes — but it amounts to a radical grandmother cell hypothesis. This type of semantic atomism strains credulity (see, for example, REF.[3]). Alternatively, combinations of semantic contents carried by individual nodes might give rise to complex contents. One node represents snow, another white and a third the predication relation (the 'is' in 'snow is white'). But, then, complex contents are not in fact represented at all, because the only carriers of semantic content in the Sherringtonian view are individual nodes. Although not false in principle, such an account flies in the face of numerous aspects of human psychology. Take, for example, language. Sentences are made up of words, but the meaning of the sentence extends beyond the meanings of the words. But in the Sherringtonian view, such extended meanings are not represented in the brain. That is wildly implausible.

Besides these representational shortcomings, the computational resources of the Sherringtonian view are also too constrained. To perform computations, the Sherringtonian view has only two elements at the algorithmic level: nodes and connections. The Sherringtonian view restricts computation to systems that satisfy node to node descriptions. To date, no one has yet described a way to execute computations by fast reconfigurations in the connections between nodes. Hence, the computations must be able to be executed by the transfer functions in the individual nodes. But this limits the types of computation that the system can perform. First, distributed representations must be analysed into contents that can be carried and transformed by single nodes. Second, if there are only local representations, then computation cannot be the result of the action and interaction of multiple components. Both of these constraints limit the types of computation available to the Sherringtonian view.

Computation and Hopfieldianism. The Hopfieldian view, by contrast, is more flexible both representationally and computationally — so much so that its computational descriptions subsume those of the Sherringtonian. A cognitive system whose operations and transformation truly occur at the level of neuron to neuron descriptions can also be described as one that moves through a neural space characterized in basic dynamical terms. The converse, however, does not hold; there may be cognitive systems that do not satisfy such point to point descriptions. As a result, the Hopfieldian view covers more systems.

Hopfieldianism characterizes cognition through identification of new first-level explainers that take the form of an object latent in neural activity described by variation in a low-dimensional projection of a high-dimensional neural state space. Indeed, mass or aggregate action in the nervous system, especially in multimodal cortex, may be so dominant as to obviate the need for a Sherringtonian computational framework with its focus on node to node architectures. As a result, the actions of these objects are invariant to swapping in or out different neurons. Neural spiking might even turn out to be epiphenomenal even if spikes correlate with the actual representation.

Neurofunctional spaces. These new first-level explanatory objects possess both functional and neural properties (neurofunctional properties) (FIG. 4). An example of this model of explanation is provided by recent work on movement planning. Planning can be explained at a neurofunctional level as a trajectory through a neural state space that culminates in the initialization of a second dynamical system for movement execution[110–113]. Movements to a target are coded by trajectories through the state space, with similar movements

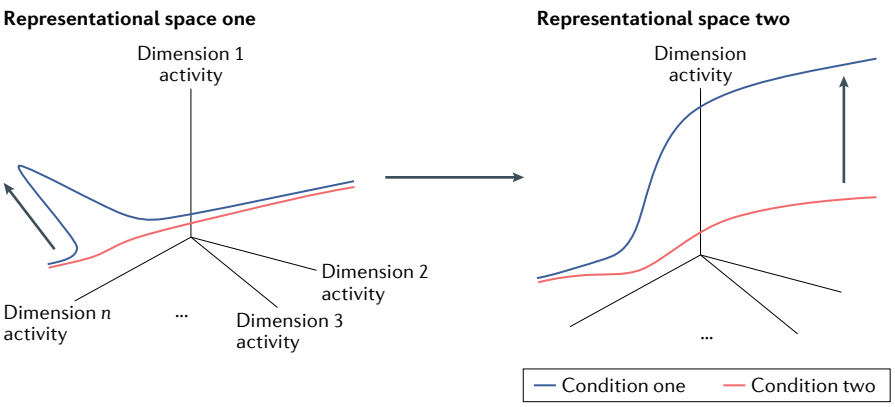


Fig. 4 | **Illustration of the revolution in Hopfieldian algorithmic approaches.** The Hopfieldian model for the algorithmic level consists of activity spaces and metric transformations between them. Representations are basins in those spaces and computations are transformations between spaces. Here, activity in one high-dimensional representational space in condition one has an excursion along some dimension. This excursion is absent in condition two. The second representational space contains a dimension that represents the integration of activity along the excursion dimension in space one. Following the excursion in condition one, the second representational space is shifted along that dimension in comparison with condition two. As a result, representational space two presents a bifurcation in the neural trajectories, a sort of neural behaviour that both plays an explanatory role in cognition and itself needs to be explained as a piece of cognitive behaviour. The set of neural trajectories traces out a lower-dimensional neural manifold that cannot be explained by the Sherringtonian approach in terms of nodes and connections. Further, the computational role for the bifurcation needs to be described.

evoking similar trajectories before a go cue. The convergence of these trajectories corresponds to planning through the setting of movement goals such as target location, trajectory shape and other features of the equivalence class of movements that execute the goal[114]. These trajectories are a form of behaviour; neurofunctional invariants across different tasks and contexts that need themselves to be explained. Notably, these trajectories in state space are representational in our more full-blooded sense because the same representations occur in different sensory contexts, are used by the system for the control of movement and are sensitive to errors and other types of accuracy-related computation. This is the formation of a plan because movement execution itself is controlled by rotational dynamics in another system that is initialized by these planning dynamics[111]. In this account, a representational entity instructs a non-representational one.

This style of explanation generalizes to other psychological explanations that feature these new neurally derived first-level explainers[115]. Representationally, different semantic contents are assigned to different regions of the latent state space, and these state spaces contribute their contents to a range of cognitive phenomena, again consistent with our concept of representation. Economic decisions may require the untangling of value information with the use of high-dimensional neural representations to form value representations[116]. Successful recall requires enough representational dimensions latent in neural populations to cover the task demands[78]. Ease of learning depends on the location of neural activity relative to a low-dimensional manifold embedded in high-dimensional neural space[117]. Cortical trajectories are used in the estimation of temporal intervals[118,119] and, in particular, different subspaces of variability in low-dimensional projections from high-dimensional neural spaces represent different intervals in the frontal cortex[120]. In a Bayesian temporal estimation task, the curvature of the lower-dimensional manifold represents priors, an explicit representational role assigned to a dynamical feature in the state space that simply cannot be captured by Sherringtonian means[121]. The same lessons are apparent from investigations of neural computation. For example, the prefrontal cortex plays a role in context-dependent perceptual decisions by movement through a task-defined representational space[75]. Working memory relies on routing trajectories through a low-dimensional manifold[122]. Rule-based reasoning and switching relies on population-level differentiation of evidence for rules in the anterior cingulate cortex[123]. This diverse set of representational and computational phenomena illustrate the explanatory power of the Hopfieldian view.

This ability to have a first-order explainer of a hybrid kind that provides understanding is a great benefit of the Hopfieldian approach. Cognitive phenomena are explained by decomposing each cognitive function into its subfunctions along with a visualization of their neural dynamics as trajectories through state spaces[103]. These depictions are analogous to Feynman diagrams, which provide an intuitive visualization of the interactions of subatomic particles that are otherwise complex and difficult to understand[124]. Such understanding has been proposed as essential to scientific explanation[125]. By contrast, recent attempts to make similar explanatory claims for neural network models that rely on connectivity metrics simply fall back on the usual shortcomings of the Sherringtonian approach: providing either opaque quantitative descriptions or testing models that themselves are couched in terms of connections without reference to either representations or computations[126].

Granted the explanatory power of the Hopfieldian view, what are the next steps in the development of the Hopfieldian research programme? Specific recommendations will depend on the particulars of the cognitive phenomena being investigated. However, some general recommendations can be made. First, neural data are no longer relegated to the role of mere implementational detail. Rather, neural activity can be a rich source of information along with behaviour to construct theories of cognition. This is a different way of thinking of neural data that emphasizes the discovery of cognitive computation directly from neural activity embedded in low-dimensional manifolds.

Second, there is a need for quantitative and conceptual advances to realize the promise of the Hopfieldian approach. Trajectories in low-dimensional state spaces are clearly operations over representations. But what are those operations? How can they be mathematically characterized? How do those mathematical characterizations relate to the problems facing the organism that are described at the ecological level (BOX 2)? A new theory of computation via these neural objects is needed. Further, computation is the transformation of representations. But those representations are also low-dimensional neural objects. How do state space transformations operate on representations that are also low-dimensional neural objects? Simply describing these operations in terms of linear dynamical systems, for example, is insufficient to connect the behaviour of these states spaces to their ecological function.

Third, Hopfieldian systems need to interact with Sherringtonian ones (see REF.[127]). Sherringtonian circuits are obviously relevant at the input and output ends of complex nervous systems, such as in sense organs or in the spinal cord. The information operated on by Hopfieldian circuits enters the system by way of Sherringtonian-type node and connection processes, and the influence of Hopfieldian circuits on behaviour must pass through the final common path and the Sherringtonian-type node and connection circuits that transmit signals from the cortex to the body and that interact with the skeletomuscular system. But how does such interaction occur? A theory of the interface between the two types of circuit is needed to integrate representational computations

Glossary

Content
The referent of a state, what the state is about.

Dimensionality
The set of basis elements whose combinations can describe any point in that space.

Exclusive disjunction
Either A or B but not both A and B.

Neural spaces
Conceptualizations of brain regions as *N*-dimensional spaces where each *N*th dimension is a representation of a neuron and the value along the dimension is the firing rate of that neuron.

Perceptrons
Early artificial neural network models.

Reticularism
An early idea about the brain's biological organization that maintained the brain is a continuous network not divisible into cells.

Semantic representations
Representations that have semantic content and can be mapped on to the content given some context of use.

State
A point or a region of neural space.

Tonotopy
An orderly arrangement of the representation of auditory tones in the brain from lowest to highest.

with the input and output ends of cognitive systems.

The Hopfieldian view may itself conceal a deeper truth about cognitive function. The brain may solve challenges at the ecological level by putting together computations from the dynamics of subcellular, neuronal, small circuit or large population components. Thus, first-level explanations of cognition may require a novel neuroscience that constructs state spaces across many spatio-temporal scales from the single molecule to the whole brain. The resultant dynamical objects of this novel cognitive neuroscience along with their functional interpretations would be the first-level explainers for cognition. The implementation-level details of the variety of neural phenomena would be the second-level explainers; that is, they would explain the novel objects and only indirectly the cognitive phenomena. Thus, Hopfieldianism, as described here for aggregate neural population data, may just be one of a wider range of ways that a theorist can construct new dynamical objects to help understand cognition. This has important implications for AI because whereas it might be possible to abstract the properties of neural populations, other tissue properties may also be important in the construction of these novel explanatory objects. The theorist should then cast a wide representational and computational net, utilizing evidence at all levels from both brain and behaviour to draw conclusions about the nature of the mind.

David L. Barack[1,2]✉ *and John W. Krakauer*[3,4,5,6]✉

[1]*Department of Philosopy, University of Pennsylvania, Philadelphia, PA, USA.*

[2]*Department of Neuroscience, University of Pennsylvania, Philadelphia, PA, USA.*

[3]*Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA.*

[4]*Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, MD, USA.*

[5]*Department of Physical Medicine and Rehabilitation, Johns Hopkins University School of Medicine, Baltimore, MD, USA.*

[6]*The Santa Fe Institute, Santa Fe, NM, USA.*

✉*e-mail: dbarack@gmail.com; jkrakau1@jhmi.edu*

https://doi.org/10.1038/s41583-021-00448-6
Published online: 15 April 2021

1. Gallistel, C. R. & King, A. P. *Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience* Vol. 3 (Wiley, 2009).
2. Goodman, N. *Languages of Art: An Approach to a Theory of Symbols* (Hackett Publishing, 1976).
3. Fodor, J. A. Propositional attitudes. *Monist* **61**, 501–523 (1978).
4. Fodor, J. A. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (MIT Press, 1987).
5. Fodor, J. A. *A Theory of Content and Other Essays* (MIT Press, 1990).
6. Cummins, R. *Meaning and Mental Representation* (MIT Press, 1989).
7. Cummins, R., Putnam, H. & Block, N. *Representations, Targets, and Attitudes* (MIT Press, 1996).
8. Millikan, R. G. *Language, Thought, and Other Biological Categories: New Foundations for Realism* (MIT Press, 1984).
9. Ramsey, W. M. *Representation Reconsidered* (Cambridge Univ. Press, 2007).
10. Shea, N. *Representation in Cognitive Science* (Oxford Univ. Press, 2018).
11. Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356 (2016).
12. Rajalingham, R. & DiCarlo, J. J. Reversible inactivation of different millimeter-scale regions of primate IT results in different patterns of core object recognition deficits. *Neuron* **102**, 493–505 (2019).
13. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
14. Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B. & Tootell, R. B. H. Faces and objects in macaque cerebral cortex. *Nat. Neurosci.* **6**, 989–995 (2003).
15. Bao, P., She, L., McGill, M. & Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
16. Kriegeskorte, N. & Diedrichsen, J. Peeling the onion of brain representations. *Annu. Rev. Neurosci.* **42**, 407–432 (2019).
17. Yuste, R. From the neuron doctrine to neural networks. *Nat. Rev. Neurosci.* **16**, 487–497 (2015).
18. Eichenbaum, H. Barlow versus Hebb: when is it time to abandon the notion of feature detectors and adopt the cell assembly as the unit of cognition? *Neurosci. Lett.* **680**, 88–93 (2018).
19. Sherrington, C. S. Observations on the scratch-reflex in the spinal dog. *J. Physiol.* **34**, 1–50 (1906).
20. Barlow, H. B. Summation and inhibition in the frog's retina. *J. Physiol.* **119**, 69–88 (1953).
21. Parker, D. Complexities and uncertainties of neuronal network function. *Philos. Trans. R. Soc. B Biol. Sci.* **361**, 81–99 (2006).
22. Tye, K. M. & Uchida, N. Editorial overview: Neurobiology of behavior. *Curr. Opin. Neurobiol.* **49**, iv–ix (2020).
23. Marder, E., Goeritz, M. L. & Otopalik, A. G. Robust circuit rhythms in small circuits arise from variable circuit components and mechanisms. *Curr. Opin. Neurobiol.* **31**, 156–163 (2015).
24. Creutzfeldt, O. D. Generality of the functional structure of the neocortex. *Naturwissenschaften* **64**, 507–517 (1977).
25. Douglas, R. J., Martin, K. A. & Whitteridge, D. A canonical microcircuit for neocortex. *Neural Comput.* **1**, 480–488 (1989).
26. Harris, K. D. & Shepherd, G. M. The neocortical circuit: themes and variations. *Nat. Neurosci.* **18**, 170–181 (2015).
27. Britten, K. H., Shadlen, M. N., Newsome, W. T. & Movshon, J. A. Responses of neurons in macaque MT to stochastic motion signals. *Vis. Neurosci.* **10**, 1157–1169 (1993).
28. Salzman, C. D. & Newsome, W. T. Neural mechanisms for forming a perceptual decision. *Science* **264**, 231–237 (1994).
29. Roitman, J. D. & Shadlen, M. N. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci.* **22**, 9475–9489 (2002).
30. Gold, J. I. & Shadlen, M. N. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron* **36**, 299–308 (2002).
31. Shadlen, M. & Newsome, W. Motion perception: seeing and deciding. *Proc. Natl Acad. Sci. USA* **93**, 628–633 (1996).
32. Mazurek, M. E., Roitman, J. D., Ditterich, J. & Shadlen, M. N. A role for neural integrators in perceptual decision making. *Cereb. Cortex* **13**, 1257–1269 (2003).
33. Ditterich, J. Stochastic models of decisions about motion direction: behavior and physiology. *Neural Netw.* **19**, 981–1012 (2006).
34. Zeki, S. M. Cells responding to changing image size and disparity in the cortex of the rhesus monkey. *J. Physiol.* **242**, 827–841 (1974).
35. Britten, K. H., Shadlen, M. N., Newsome, W. T. & Movshon, J. A. Responses of neurons in macaque MT to stochastic motion signals. *Vis. Neurosci.* **10**, 1157–1169 (1993).
36. Britten, K. H., Shadlen, M. N., Newsome, W. T. & Movshon, J. A. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J. Neurosci.* **12**, 4745–4765 (1992).
37. Blatt, G. J., Andersen, R. A. & Stoner, G. R. Visual receptive field organization and cortico-cortical connections of the lateral intraparietal area (area LIP) in the macaque. *J. Comp. Neurol.* **299**, 421–445 (1990).
38. Latimer, K. W., Yates, J. L., Meister, M. L., Huk, A. C. & Pillow, J. W. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science* **349**, 184–187 (2015).
39. Katz, L. N., Yates, J. L., Pillow, J. W. & Huk, A. C. Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature* **535**, 285–288 (2016).
40. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).
41. Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**, 386 (1958).
42. Minsky, M. & Papert, S. A. *Perceptrons: An Introduction to Computational Geometry* (MIT Press, 1969).
43. Rigotti, M., Rubin, D. B., Wang, X. J. & Fusi, S. Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. *Front. Comput. Neurosci.* **4**, 24 (2010).
44. Zador, A. M., Claiborne, B. J. & Brown, T. H. in *Advances in Neural Information Processing Systems* 51–58 (NIPS, 1991).
45. Legenstein, R. & Maass, W. Branch-specific plasticity enables self-organization of nonlinear computation in single neurons. *J. Neurosci.* **31**, 10787–10802 (2011).
46. Kimura, R. et al. Hippocampal polysynaptic computation. *J. Neurosci.* **31**, 13168–13179 (2011).
47. Bianchi, D. et al. On the mechanisms underlying the depolarization block in the spiking dynamics of CA1 pyramidal neurons. *J. Comput. Neurosci.* **33**, 207–225 (2012).
48. Gidon, A. et al. Dendritic action potentials and computation in human layer 2/3 cortical neurons. *Science* **367**, 83–87 (2020).
49. Wang, X. J. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* **36**, 955–968 (2002).
50. Hebb, D. *The Organization of Behavior* (Wiley, 1949).
51. McCrea, D. A. & Rybak, I. A. Organization of mammalian locomotor rhythm and pattern generation. *Brain Res. Rev.* **57**, 134–146 (2008).
52. Fries, P. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends Cognit. Sci.* **9**, 474–480 (2005).
53. 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).
54. Jazayeri, M. & Afraz, A. Navigating the neural space in search of the neural code. *Neuron* **93**, 1003–1014 (2017).
55. Paninski, L. & Cunningham, J. P. Neural data science: accelerating the experiment–analysis–theory cycle in large-scale neuroscience. *Curr. Opin. Neurobiol.* **50**, 232–241 (2018).
56. Hu, Y., Trousdale, J., Josić, K. & Shea-Brown, E. Motif statistics and spike correlations in neuronal networks. *J. Stat. Mech. Theory Exp.* **2013**, P03012 (2013).
57. Hu, Y., Trousdale, J., Josić, K. & Shea-Brown, E. Local paths to global coherence: cutting networks down to size. *Phys. Rev. E* **89**, 032802 (2014).
58. Hu, Y. et al. Feedback through graph motifs relates structure and function in complex networks. *Phys. Rev. E* **98**, 062312 (2018).
59. Recanatesi, S., Ocker, G. K., Buice, M. A. & Shea-Brown, E. Dimensionality in recurrent spiking networks: global trends in activity and local origins in connectivity. *PLoS Comput. Biol.* **15**, e1006446 (2019).
60. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl Acad. Sci. USA* **79**, 2554–2558 (1982).
61. Hopfield, J. J. Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Natl Acad. Sci. USA* **81**, 3088–3092 (1984).
62. Hopfield, J. J. & Tank, D. W. Computing with neural circuits: a model. *Science* **233**, 625–633 (1986).
63. Ramón y Cajal, S. Estudios sobre la corteza cerebral humana. Corteza visual. *Rev. Trim. Microgr.* **4**, 1–63 (1899).

64. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133 (1943).
65. Buzsaki, G. Neural syntax: cell assemblies, synapsembles, and readers. *Neuron* **68**, 362–385 (2010).
66. Saxena, S. & Cunningham, J. P. Towards the neural population doctrine. *Curr. Opin. Neurobiol.* **55**, 103–111 (2019).
67. Mesulam, M.-M. From sensation to cognition. *Brain J. Neurol.* **121**, 1013–1052 (1998).
68. Anderson, M. L. *After Phrenology* (Oxford Univ. Press, 2014).
69. Sporns, O. *Networks of the Brain* (MIT Press, 2010).
70. Lashley, K. S. Mass action in cerebral function. *Science* **73**, 245–254 (1931).
71. Trautmann, E. M. et al. Accurate estimation of neural population dynamics without spike sorting. *Neuron* **103**, 292–308 (2019).
72. Clark, A. *A Theory of Sentience* (Clarendon Press, 2000).
73. Gärdenfors, P. *Conceptual Spaces: The Geometry of Thought* (MIT Press, 2004).
74. Meister, M. L., Hennig, J. A. & Huk, A. C. Signal multiplexing and single-neuron computations in lateral intraparietal area during decision-making. *J. Neurosci.* **33**, 2254–2267 (2013).
75. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
76. Warden, M. R. & Miller, E. K. The representation of multiple objects in prefrontal neuronal delay activity. *Cereb. Cortex* **17**, i41–i50 (2007).
77. Warden, M. R. & Miller, E. K. Task-dependent changes in short-term memory in the prefrontal cortex. *J. Neurosci.* **30**, 15801–15810 (2010).
78. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
79. Moser, E. I. et al. Grid cells and cortical representation. *Nat. Rev. Neurosci.* **15**, 466 (2014).
80. Moser, E. I., Moser, M.-B. & McNaughton, B. L. Spatial representation in the hippocampal formation: a history. *Nat. Neurosci.* **20**, 1448 (2017).
81. Fyhn, M., Molden, S., Witter, M. P., Moser, E. I. & Moser, M.-B. Spatial representation in the entorhinal cortex. *Science* **305**, 1258–1264 (2004).
82. Hafting, T., Fyhn, M., Molden, S., Moser, M.-B. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**, 801–806 (2005).
83. Sargolini, F. et al. Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science* **312**, 758–762 (2006).
84. Tolman, E. C. Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189 (1948).
85. Constantinescu, A. O., O'Reilly, J. X. & Behrens, T. E. Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468 (2016).
86. Behrens, T. E. et al. What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* **100**, 490–509 (2018).
87. Sorscher, B., Mel, G., Ganguli, S. & Ocko, S. in *Advances in Neural Information Processing Systems* 10003–10013 (NeurIIPS, 2019).
88. Cueva, C. J. & Wei, X.-X. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. Preprint at https://arxiv.org/abs/1803.07770 (2018).
89. Banino, A. et al. Vector-based navigation using grid-like representations in artificial agents. *Nature* **557**, 429–433 (2018).
90. Felleman, D. & Van Essen, D. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
91. Zador, A. M. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat. Commun.* **10**, 1–7 (2019).
92. Polger, T. W. & Shapiro, L. A. *The Multiple Realization Book* (Oxford Univ. Press, 2016).
93. Bechtel, W. A bridge between cognitive science and neuroscience: the functional architecture of mind. *Philos. Stud.* **44**, 319–330 (1983).
94. Pylyshyn, Z. W. *Computation and Cognition* (Cambridge Univ. Press, 1984).
95. Ramon y Cajal, S. *Estructura de los centros nerviosos de las aves* [Spanish] (1888).
96. Sherrington, C. *The Integrative Action of the Central Nervous System* (Archibald Constable, 1906).
97. Barlow, H. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* **1**, 371–394 (1972).
98. Martin, K. A. A brief history of the "feature detector". *Cereb. Cortex* **4**, 1–7 (1994).
99. Shepherd, G. M. *Foundations of the Neuron Doctrine* (Oxford Univ. Press, 2015).
100. Kuhn, T. S. *The Structure of Scientific Revolutions* (Univ. of Chicago Press, 1962).
101. Haberkern, H. & Jayaraman, V. Studying small brains to understand the building blocks of cognition. *Curr. Opin. Neurobiol.* **37**, 59–65 (2016).
102. Cobb, M. *The Idea of the Brain: The Past and Future of Neuroscience* (Basic Books, 2020).
103. Barack, D. L. Mental machines. *Biol. Philos.* **34**, 63 (2019).
104. Fuster, J. *The Prefrontal Cortex* (Academic Press, 2008).
105. Arbib, M. A., Plangprasopchok, A., Bonaiuto, J. & Schuler, R. E. A neuroinformatics of brain modeling and its implementation in the Brain Operation Database BODB. *Neuroinformatics* **12**, 5–26 (2014).
106. Carandini, M. & Heeger, D. J. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2012).
107. Wong, K.-F. & Wang, X.-J. A recurrent network mechanism of time integration in perceptual decisions. *J. Neurosci.* **26**, 1314–1328 (2006).
108. Song, H. F., Yang, G. R. & Wang, X.-J. Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife* **6**, e21492 (2017).
109. Song, H. F., Yang, G. R. & Wang, X.-J. Training excitatory–inhibitory recurrent neural networks for cognitive tasks: a simple and flexible framework. *PLoS Comput. Biol.* **12**, e1004792–e1004792 (2016).
110. Churchland, M. M., Byron, M. Y., Ryu, S. I., Santhanam, G. & Shenoy, K. V. Neural variability in premotor cortex provides a signature of motor preparation. *J. Neurosci.* **26**, 3697–3712 (2006).
111. Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Ryu, S. I. & Shenoy, K. V. Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron* **68**, 387–400 (2010).
112. Wong, A. L., Haith, A. M. & Krakauer, J. W. Motor planning. *Neuroscientist* **21**, 385–398 (2015).
113. Haith, A. M. & Bestmann, S. in *The Cognitive Neurosciences VI* (eds Poeppel, D., Mangun, R., & Gazzaniga, M. S.) 541–548 (MIT Press, 2020).
114. Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical control of arm movements: a dynamical systems perspective. *Annu. Rev. Neurosci.* **36**, 337–359 (2013).
115. Vyas, S., Golub, M. D., Sussillo, D. & Shenoy, K. Computation through neural population dynamics. *Annu. Rev. Neurosci.* **43**, 249–275 (2020).
116. Yoo, S. B. M. & Hayden, B. Y. Economic choice as an untangling of options into actions. *Neuron* **99**, 434–447 (2018).
117. Golub, M. D. et al. Learning by neural reassociation. *Nat. Neurosci.* **21**, 607–616 (2018).
118. Wang, J., Narain, D., Hosseini, E. A. & Jazayeri, M. Flexible timing by temporal scaling of cortical responses. *Nat. Neurosci.* **21**, 102–110 (2018).
119. Egger, S. W., Le, N. M. & Jazayeri, M. A neural circuit model for human sensorimotor timing. *Nat. Commun.* **11**, 3933 (2020).
120. 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).
121. Sohn, H., Narain, D., Meirhaeghe, N. & Jazayeri, M. Bayesian computation through cortical latent dynamics. *Neuron* **103**, 934–947 (2019).
122. 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 (2017).
123. Sarafyazd, M. & Jazayeri, M. Hierarchical reasoning by neural circuits in the frontal cortex. *Science* **364**, eaav8911 (2019).
124. Feynman, R. P. Space–time approach to quantum electrodynamics. *Phys. Rev.* **76**, 769 (1949).
125. De Regt, H. W. *Understanding Scientific Understanding* (Oxford Univ. Press, 2017).
126. Bertolero, M. A. & Bassett, D. S. On the nature of explanations offered by network science: A perspective from and for practicing neuroscientists. *Top. Cogn. Sci.* **12**, 1272–1293 (2020).
127. Kohn, A. et al. Principles of corticocortical communication: proposed schemes and design considerations. *Trends Neurosci.* **43**, 725–737 (2020).
128. Nelson, S. B. Cortical microcircuits: diverse or canonical? *Neuron* **36**, 19–27 (2002).
129. Churchland, P. M. Cognitive neurobiology: a computational hypothesis for laminar cortex. *Biol. Philos.* **1**, 25–51 (1986).
130. Lisman, J. et al. The molecular basis of CaMKII function in synaptic and behavioural memory. *Nat. Rev. Neurosci.* **3**, 175–190 (2002).
131. Janak, P. H. & Tye, K. M. From circuits to behaviour in the amygdala. *Nature* **517**, 284–292 (2015).
132. Churchland, M. M. et al. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nat. Neurosci.* **13**, 369–378 (2010).
133. Marr, D. *Vision* (Henry Holt, 1982).
134. Sterelny, K. *The Representational Theory of Mind: An Introduction* (Blackwell, 1990).
135. Shagrir, O. Marr on computational-level theories. *Philos. Sci.* **77**, 477–500 (2010).
136. Haugeland, J. *Artificial Intelligence: The Very Idea* (MIT Press, 1985).

Author contributions
D.L.B. and J.W.K. contributed equally to this work.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Neuroscience thanks T. Behrens, who co-reviewed with A. Baram; R. Krauzlis; and E. Miller for their contribution to the peer review of this work.