REVIEW ARTICLE
https://doi.org/10.1038/s41593-018-0210-5

nature neuroscience

Cognitive computational neuroscience

Nikolaus Kriegeskorte[1]* and Pamela K. Douglas[2]

To learn how cognition is implemented in the brain, we must build computational models that can perform cognitive tasks, and test such models with brain and behavioral experiments. Cognitive science has developed computational models that decompose cognition into functional components. Computational neuroscience has modeled how interacting neurons can implement elementary components of cognition. It is time to assemble the pieces of the puzzle of brain computation and to better integrate these separate disciplines. Modern technologies enable us to measure and manipulate brain activity in unprecedentedly rich ways in animals and humans. However, experiments will yield theoretical insight only when employed to test brain-computational models. Here we review recent work in the intersection of cognitive science, computational neuroscience and artificial intelligence. Computational models that mimic brain information processing during perceptual, cognitive and control tasks are beginning to be developed and tested with brain and behavioral data.

Understanding brain information processing requires that we build computational models that are capable of performing cognitive tasks. The argument in favor of task-performing computational models was well articulated by Allen Newell in 1973 in his commentary "You can't play 20 questions with nature and win"[1]. Newell was criticizing the state of cognitive psychology. The field was in the habit of testing one hypothesis about cognition at a time, in the hope that forcing nature to answer a series of binary questions would eventually reveal the brain's algorithms. Newell argued that testing verbally defined hypotheses about cognition might never lead to a computational understanding. Hypothesis testing, in his view, needed to be complemented by the construction of comprehensive task-performing computational models. Only synthesis in a computer simulation can reveal what the interaction of the proposed component mechanisms actually entails and whether it can account for the cognitive function in question. If we did have a full understanding of an information-processing mechanism, then we should be able to engineer it. "What I cannot create, I do not understand," in the words of physicist Richard Feynman, who left this sentence on his blackboard when he died in 1988.

Here we argue that task-performing computational models that explain how cognition arises from neurobiologically plausible dynamic components will be central to a new cognitive computational neuroscience. We first briefly trace the steps of the cognitive and brain sciences and then review several exciting recent developments that suggest that it might be possible to meet the combined ambitions of cognitive science (to explain how humans learn and think)[2] and computational neuroscience (to explain how brains adapt and compute)[3] using neurobiologically plausible artificial intelligence (AI) models.

In the spirit of Newell's critique, the transition from cognitive psychology to cognitive science was defined by the introduction of task-performing computational models. Cognitive scientists knew that understanding cognition required AI and brought engineering to cognitive studies. In the 1980s, cognitive science made important advances with symbolic cognitive architectures[4,5] and neural networks[6], using human behavioral data to adjudicate between candidate computational models. However, computer hardware and machine learning were not sufficiently advanced to simulate cognitive processes in their full complexity. Moreover, these early developments relied on behavioral data alone and did not leverage constraints provided by the anatomy and activity of the brain.

With the advent of human functional brain imaging, scientists began to relate cognitive theories to the human brain. This endeavor came to be called cognitive neuroscience[7]. Cognitive neuroscientists began by mapping cognitive psychology's boxes (information-processing modules) and arrows (interactions between modules) onto the brain. This was a step forward in terms of engaging brain activity, but a step back in terms of computational rigor. Methods for testing the task-performing computational models of cognitive science with brain-activity data had not been conceived. As a result, cognitive science and cognitive neuroscience parted ways in the 1990s.

Cognitive psychology's tasks and theories of high-level functional modules provided a reasonable starting point for mapping the coarse-scale organization of the human brain with functional imaging techniques, including electroencephalography, positron emission tomography and early functional magnetic resonance imaging (fMRI), which had low spatial resolution. Inspired by cognitive psychology's notion of the module[8], cognitive neuroscience developed its own game of 20 questions with nature. A given study would ask whether a particular cognitive module could be found in the brain. The field mapped an ever increasing array of cognitive functions to brain regions, providing a useful rough draft of the global functional layout of the human brain.

A brain map, at whatever scale, does not reveal the computational mechanism (Fig. 1). However, mapping does provide constraints for theory. After all, information exchange incurs costs that scale with the distance between the communicating regions—costs in terms of physical connections, energy and signal latency. Component placement is likely to reflect these costs. We expect regions that need to interact at high bandwidth and short latency to be placed close together[9]. More generally, the topology and geometry of a biological neural network constrain its dynamics, and thus its functional mechanism. Functional localization results, especially in combination with anatomical connectivity, may therefore ultimately prove useful for modeling brain information processing.

[1]Department of Psychology, Department of Neuroscience, Department of Electrical Engineering, Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA. [2]Center for Cognitive Neuroscience, University of California, Los Angeles, Los Angeles, CA, USA.
*e-mail: n.kriegeskorte@columbia.edu

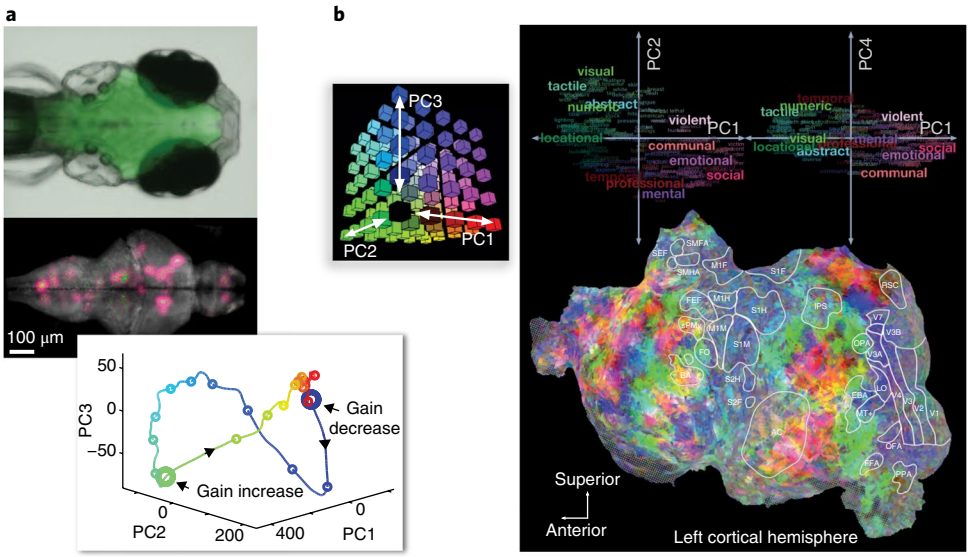


Fig. 1 | Modern imaging techniques provide unprecedentedly detailed information about brain activity, but data-driven analyses support only limited insights. a, Two-photon calcium imaging results[121] show single-neuron activity for a large population of cells measured simultaneously in larval zebrafish while the animals interact with a virtual environment. **b**, Human fMRI results[70] reveal a detailed map of semantically selective responses while a subject listened to a story. These studies illustrate, on the one hand, the power of modern brain-activity measurement techniques at different scales (**a**,**b**) and, on the other, the challenge of drawing insights about brain computation from such datasets. Both studies measured brain activity during complex, time-continuous, naturalistic experience and used principal component analysis (**a**, bottom; **b**, top) to provide an overall view of the activity patterns and their representational significance. PC, principal component.

Despite methodological challenges[10,11], many of the findings of cognitive neuroscience provide a solid basis on which to build. For example, the findings of face-selective regions in the human ventral stream[12] have been thoroughly replicated and generalized. Nonhuman primates probed with fMRI exhibit similar face-selective regions, which had evaded explorations with invasive electrodes because the latter do not provide continuous images over large fields of view. Localized with fMRI and probed with invasive electrode recordings, the primate face patches revealed high densities of face-selective neurons[13], with invariances emerging at higher stages of hierarchical processing, including mirror-symmetric tuning and view-tolerant representations of individual faces in the anterior-most patch[14]. The example of face perception illustrates, on one hand, the solid progress in mapping the anatomical substrate and characterizing neuronal responses[15] and, on the other, the lack of definitive computational models. The literature does provide clues to the computational mechanism. A brain-computational model of face recognition[16] will have to explain the spatial clusters of face-selective units and the selectivities and invariances observed with fMRI[17,18] and invasive recordings[14,19].

Cognitive neuroscience has mapped the global functional layout of the human and nonhuman primate brain[20]. However, it has not achieved a full computational account of brain information processing. The challenge ahead is to build computational models of brain information processing that are consistent with brain structure and function and perform complex cognitive tasks. The following recent developments in cognitive science, computational neuroscience and artificial intelligence suggest that this may be achievable.

1. Cognitive science has proceeded from the top down, decomposing complex cognitive processes into their computational components. Unencumbered by the need to make sense of brain data, it has developed task-performing computational models at the cognitive level. One success story is that of Bayesian cognitive models, which optimally combine prior knowledge about the world with sensory evidence[21–23]. Initially applied to basic sensory and motor processes[23,24], Bayesian models have begun to engage complex cognition, including the way our minds model the physical and social world[2]. These developments occurred in interaction with statistics and machine learning, where a unified perspective on probabilistic empirical inference has emerged. This literature provides essential computational theory for understanding the brain. In addition, it provides algorithms for approximate inference on generative models that can grow in complexity with the available data—as might be required for real-world intelligence[25,26].

2. Computational neuroscience has taken a bottom-up approach, demonstrating how dynamic interactions between biological neurons can implement computational component functions. In the past two decades, the field developed mathematical models of elementary computational components and their implementation with biological neurons[27,28]. These include components for sensory coding[29,30], normalization[31], working memory[32], evidence accumulation and decision mechanisms[33–35], and motor control[36]. Most of these component functions are computationally simple, but they provide building blocks for cognition. Computational neuroscience has also begun to test complex computational models that can explain high-level sensory and cognitive brain representations[37,38].

3. Artificial intelligence has shown how component functions can be combined to create intelligent behavior. Early AI failed to live up to its promise because the rich world knowledge required for feats of intelligence could not be either engineered or automatically learned. Recent advances in machine learning, boosted by growing computational power and larger datasets from which to learn, have brought progress at perceptual[39], cognitive[40] and control challenges[41]. Many advances were driven by cognitive-level symbolic models. Some of the most important recent advances are driven by deep neural network models, composed of units that compute linear combinations of their inputs, followed by static nonlinearities[42]. These models employ only a small subset of the dynamic capabilities of biological neurons, abstracting from fundamental features such as action potentials. However, their functionality is inspired by brains and could be implemented with biological neurons.

The three disciplines contribute complementary elements to biologically plausible computational models that perform cognitive tasks and explain brain information processing and behavior (Fig. 2). Here we review the first steps in the literature toward a cognitive computational neuroscience that meets the combined criteria for success of cognitive science (computational models that perform cognitive tasks and explain behavior) and computational neuroscience (neurobiologically plausible mechanistic models that explain brain activity). If computational models are to explain animal and human cognition, they will have to perform feats of intelligence. AI, and in particular machine learning, is therefore a key discipline that provides the theoretical and technological foundation for cognitive computational neuroscience.

The overarching challenge is to build solid bridges between theory (instantiated in task-performing computational models) and experiment (providing brain and behavioral data). The first part of this review describes bottom-up developments that begin with experimental data and attempt to build bridges from the data in the direction of theory[43]. Given brain-activity data, connectivity models aim to reveal the large-scale dynamics of brain activation; decoding and encoding models aim to reveal the content and format of brain representations. The models employed in this literature provide constraints for computational theory, but they do not in general perform the cognitive tasks in question and thus fall short of explaining the computational mechanism underlying task performance.

The second part of this article describes developments that proceed in the opposite direction, building bridges from theory to experiment[37,38,44]. We review emerging work that has begun to test task-performing computational models with brain and behavioral data. The models include cognitive models, specified at an abstract computational level, whose implementation in biological brains has yet to be explained, and neural network models, which abstract from many features of neurobiology, but could plausibly be implemented with biological neurons. This emerging literature suggests the beginnings of an integrative approach to understanding brain computation, where models are required to perform cognitive tasks, biology provides the admissible component functions, and the computational mechanisms are optimized to explain detailed patterns of brain activity and behavior.

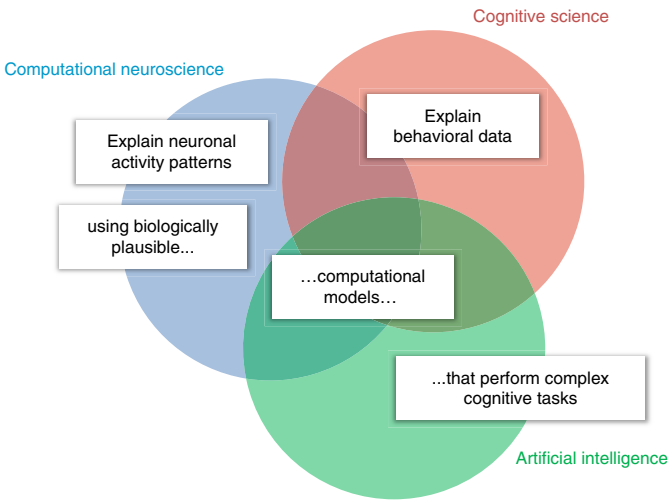


Fig. 2 | What does it mean to understand how the brain works? The goal of cognitive computational neuroscience is to explain rich measurements of neuronal activity and behavior in animals and humans by means of biologically plausible computational models that perform real-world cognitive tasks. Historically, each of the disciplines (circles) has tackled a subset of these challenges (white labels). Cognitive computational neuroscience strives to meet all the challenges simultaneously.

From experiment toward theory

Models of connectivity and dynamics. One path from measured brain activity toward a computational understanding is to model the brain's connectivity and dynamics. Connectivity models go beyond the localization of activated regions and characterize the interactions between regions. Neuronal dynamics can be measured and modeled at multiple scales, from local sets of interacting neurons to whole-brain activity[45]. A first approximation of brain dynamics is provided by the correlation matrix among the measured response time series, which characterizes the pairwise 'functional connectivity' between locations. The literature on resting-state networks has explored this approach[46], and linear decompositions of the space-time matrix, such as spatial independent component analysis, similarly capture simultaneous correlations between locations across time[47].

By thresholding the correlation matrix, the set of regions can be converted into an undirected graph and studied with graph-theoretic methods. Such analyses can reveal 'communities' (sets of strongly interconnected regions), 'hubs' (regions connected to many others) and 'rich clubs' (communities of hubs)[48]. Connectivity graphs can be derived from either anatomical or functional measurements. The anatomical connectivity matrix typically resembles the functional connectivity matrix because regions interact through anatomical pathways. However, the way anatomical connectivity generates functional connectivity is better modeled by taking local dynamics, delays, indirect interactions and noise into account[49]. From local neuronal interactions to large-scale spatiotemporal patterns spanning cortex and subcortical regions, generative models of spontaneous dynamics can be evaluated with brain-activity data.

Effective connectivity analyses take a more hypothesis-driven approach, characterizing the interactions among a small set of regions on the basis of generative models of the dynamics[50]. Whereas activation mapping maps the boxes of cognitive psychology onto brain regions, effective connectivity analyses map the arrows onto pairs of brain regions. Most work in this area has focused on characterizing interactions at the level of the overall activation of a brain region. Like the classical brain mapping approach, these analyses are based on regional-mean activation, measuring correlated fluctuations of overall regional activation rather than the information exchanged between regions.

Analyses of effective connectivity and large-scale brain dynamics go beyond generic statistical models such as the linear models used in activation and information-based brain mapping in that they are generative models: they can generate data at the level of the measurements and are models of brain dynamics. However, they do not capture the represented information and how it is processed in the brain.

Decoding models. Another path toward understanding the brain's computational mechanisms is to reveal what information is present in each brain region. Decoding can help us go beyond the notion of activation, which indicates the involvement of a region in a task, and reveal the information present in a region's population activity. When particular content is decodable from activity in a brain region, this indicates the presence of the information. To refer to the brain region as 'representing' the content adds a functional interpretation[51]: that the information serves the purpose of informing regions receiving these signals about the content. Ultimately, this interpretation needs to be substantiated by further analyses of how the information affects other regions and behavior[52–54].

Decoding has its roots in the neuronal-recording literature[27] and has become a popular tool for studying the content of representations in neuroimaging[55–59]. In the simplest case, decoding reveals which of two stimuli gave rise to a measured response pattern.

Box 1 | The many meanings of "model"

The word "model" has many meanings in the brain and behavioral sciences. *Data-analysis models* are generic statistical models that help establish relationships between measured variables. Examples include linear correlation, univariate multiple linear regression for brain mapping, and linear decoding analysis. Effective connectivity and causal-interaction models are, similarly, data-analysis models. They help us infer causal influences and interactions between brain regions. Data-analysis models can serve the purpose of testing hypotheses about relationships among variables (for example, correlation, information, causal influence). They are not models of brain information processing. A *box-and-arrow model*, by contrast, is an information-processing model in the form of labeled boxes that represent cognitive component functions and arrows that represent information flow. In cognitive psychology, such models provided useful, albeit ill-defined, sketches for theories of brain computation. A *word model*, similarly, is a sketch for a theory about brain information processing that is defined vaguely by a verbal description. While these are models of information processing, they do not perform the information processing thought to occur in the brain. An *oracle model* is a model of brain responses (often instantiated in a data-analysis model) that relies on information not available to the animal whose brain is being modeled. For example, a model of ventral temporal visual responses as a function of an abstract shape description, or as a function of category labels or continuous semantic features, constitutes an oracle model if the model is not capable of computing the shape, category or semantic features from images. An oracle model may provide a useful characterization of the information present in a region and its representational format, without specifying any theory as to how the representation is computed by the brain. A *brain-computational model (BCM)*, by contrast, is a model that mimics the brain information processing underlying the performance of some task at some level of abstraction. In visual neuroscience, for example, an *image-computable model* is a BCM of visual processing that takes image bitmaps as inputs and predicts brain activity and/or behavioral responses. Deep neural nets provide image-computable models of visual processing. However, deep neural nets trained by supervision rely on category-labeled images for training. Because labeled examples are not available (in comparable quantities) during biological development and learning, these models are BCMs of visual processing, but they are not BCMs of development and learning. *Reinforcement learning models* use environmental feedback that is more realistic in quality and can provide BCMs of learning processes. A *sensory encoding model* is a BCM of the computations that transform sensory input to some stage of internal representation. An *internal-transformation model* is a BCM of the transformation of representations between two stages of processing. A *behavioral decoding model* is a BCM of the transformation from some internal representation to a behavioral output. Note that the label BCM indicates merely that the model is intended to capture brain computations at some level of abstraction. A BCM may abstract from biological detail to an arbitrary degree, but must predict some aspect of brain activity and/or behavior. *Psychophysical models* that predict behavioral outputs from sensory input and *cognitive models* that perform cognitive tasks are BCMs formulated at a high level of description. The label BCM does not imply that the model is either plausible or consistent with empirical data. Progress is made by rejecting candidate BCMs on empirical grounds. Like microscale *biophysical models*, which capture biological processes that underlie brain computations, and macroscale *brain-dynamical and causal-interaction models*, BCMs are models of processes occurring in the brain. However, unlike the other types of process model, BCMs perform the information processing that is thought to be the function of brain dynamics. Finally, the term "model" is used to refer to models of the world employed by the brain, as in *model-based reinforcement learning* and *model-based cognition*.

The content of the representation can be the identity of a sensory stimulus (to be recognized among a set of alternative stimuli), a stimulus property (such as the orientation of a grating), an abstract variable needed for a cognitive operation, or an action[60]. When the decoder is linear, as is usually the case, the decodable information is in a format that can plausibly be read out by downstream neurons in a single step. Such information is said to be 'explicit' in the activity patterns[61].

Decoding and other types of multivariate pattern analysis have helped reveal the content of regional representations[55,56,58,59], providing evidence that brain-computational models must incorporate. However, the ability to decode particular information does not amount to a full account of the neuronal code: it doesn't specify the representational format (beyond linear decodability) or what other information might additionally be present. Most importantly, decoders do not in general constitute models of brain computation. They reveal aspects of the product, but not the process of brain computation.

Representational models. Beyond decoding, we would like to exhaustively characterize a region's representation, explaining its responses to arbitrary stimuli. A full characterization would also define to what extent any variable can be decoded. Representational models attempt to make comprehensive predictions about the representational space and therefore provide stronger constraints on the computational mechanism than decoding models[52,62].

Three types of representational model analysis have been introduced in the literature: encoding models[63–65], pattern component models[66] and representational similarity analysis[57,67,68]. These three methods all test hypotheses about the representational space, which are based on multivariate descriptions of the experimental conditions—for example, a semantic description of a set of stimuli, or the activity patterns across a layer of a neural network model that processes the stimuli[52].

In encoding models, each voxel's activity profile across stimuli is predicted as a linear combination of the features of the model. In pattern component models, the distribution of the activity profiles that characterizes the representational space is modeled as a multivariate normal distribution. In representational similarity analysis, the representational space is characterized by the representational dissimilarities of the activity patterns elicited by the stimuli.

Representational models are often defined on the basis of descriptions of the stimuli, such as labels provided by human observers[63,69,70]. In this scenario, a representational model that explains the brain responses in a given region provides, not a brain-computational account, but at least a descriptive account of the representation. Such an account can be a useful stepping-stone toward computational theory when the model generalizes to novel stimuli. Importantly, representational models also enable us to adjudicate among brain-computational models, an approach we will return to in the next section.

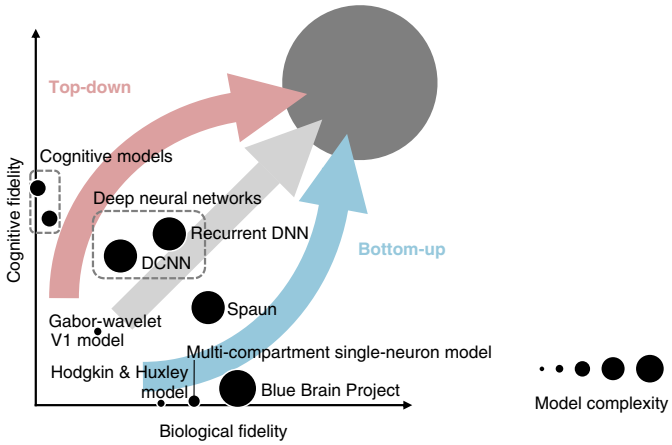


Fig. 3 | The space of process models. Models of the processes taking place in the brain can be defined at different levels of description and can vary in their parametric complexity (dot size) and in their biological (horizontal axis) and cognitive (vertical axis) fidelity. Theoreticians approach modeling with a range of primary goals. The bottom-up approach to modeling (blue arrow) aims first to capture characteristics of biological neural networks, such as action potentials and interactions among multiple compartments of single neurons. This approach disregards cognitive function so as to focus on understanding the emergent dynamics of small parts of the brain, such as cortical columns and areas, and to reproduce biological network phenomena, such as oscillations. The top-down approach (red arrow) aims first to capture cognitive functions at the algorithmic level. This approach disregards the biological implementation so as to focus on decomposing the information processing underlying task performance into its algorithmic components. The two approaches form the extremes of a continuum of paths toward the common goal of explaining how our brains give rise to our minds. Overall, there is tradeoff (negative correlation) between cognitive and biological fidelity. However, the tradeoff can turn into a synergy (positive correlation) when cognitive constraints illuminate biological function and when biology inspires models that explain cognitive feats. Because intelligence requires rich world knowledge, models of human brain information processing will have high parametric complexity (large dot in the upper right corner). Even if models that abstract from biological details can explain task performance, biologically detailed models will still be needed to explain the neurobiological implementation. This diagram is a conceptual cartoon that can help us understand the relationships between models and appreciate their complementary contributions. However, it is not based on quantitative measures of cognitive fidelity, biological fidelity and model complexity. Definitive ways to measure each of the three variables have yet to be developed. Figure inspired by ref. [122].

In this section, we considered three types of model that can help us glean computational insight from brain-activity data. Connectivity models capture aspects of the dynamic interactions between regions. Decoding models enable us to look into brain regions and reveal what might be their representational content. Representational models enable us to test explicit hypotheses that fully characterize a region's representational space. All three types of model can be used to address theoretically motivated questions—taking a hypothesis-driven approach. However, in the absence of task-performing computational models, they are subject to Newell's argument that asking a series of questions might never reveal the computational mechanism underlying the cognitive feat we are trying to explain. These methods fall short of building the bridge all the way to theory because they do not test mechanistic models that specify precisely how the information processing underlying some cognitive function might work.

From theory to experiment

To build a better bridge between experiment and theory, we first need to fully specify a theory. This can be achieved by defining the theory mathematically and implementing it in a computational model (Box 1). Computational models can reside at different levels of description, trading off cognitive fidelity against biological fidelity (Fig. 3). Models designed to capture only neuronal components and dynamics[71] tend to be unsuccessful at explaining cognitive function[72] (Fig. 3, horizontal axis). Conversely, models designed to capture only cognitive functions are difficult to relate to the brain (Fig. 3, vertical axis). To link mind and brain, models must attempt to capture aspects of both behavior and neuronal dynamics. Recent advances suggest that constraints from the brain can help explain cognitive function[42,73,74] and vice versa[37,38], turning the tradeoff into a synergy.

In this section, we focus on recent successes with task-performing models that explain cognitive functions in terms of representations and algorithms. Task-performing models have been central to psychophysics and cognitive science, where they are traditionally tested with behavioral data. An emerging literature is beginning to test task-performing models with brain-activity data as well. We will consider two broad classes of model in turn, neural network models and cognitive models.

Neural network models. Neural network models (Box 2) have a long history, with interwoven strands in multiple disciplines. In computational neuroscience, neural network models, at various levels of biological detail, have been essential to understanding dynamics in biological neural networks and elementary computational functions[27,28]. In cognitive science, they defined a new paradigm for understanding cognitive functions, called parallel distributed processing, in the 1980s[6,75], which brought the field closer to neuroscience. In AI, they have recently brought substantial advances in a number of applications[42,74], ranging from perceptual tasks (such as vision and speech recognition) to symbolic processing challenges (such as language translation), and on to motor tasks (including speech synthesis and robotic control). Neural network models provide a common language for building task-performing models that meet the combined criteria for success of the three disciplines (Fig. 2).

Like brains, neural network models can perform feedforward as well as recurrent computations[37,76]. The models driving the recent advances are deep in the sense that they comprise multiple stages of linear-nonlinear signal transformation. Models typically have millions of parameters (the connection weights), which are set so as to optimize task performance. One successful paradigm is supervised learning, wherein a desired mapping from inputs to outputs is learned from a training set of inputs (for example, images) and associated outputs (for example, category labels). However, neural network models can also be trained without supervision and can learn complex statistical structure inherent to their experiential data.

The large number of parameters creates unease among researchers who are used to simple models with small numbers of interpretable parameters. However, simple models will never enable us to explain complex feats of intelligence. The history of AI has shown that intelligence requires ample world knowledge and sufficient parametric complexity to store it. We therefore must engage complex models (Fig. 3) and the challenges they pose. One challenge is that the high parameter count renders the models difficult to understand. Because the models are entirely transparent, they can be probed cheaply with millions of input patterns to understand the internal representations, an approach sometimes called 'synthetic neurophysiology'. To address the concern of overfitting, models are evaluated in terms of their generalization performance. A vision model, for example, will be evaluated in terms of its ability to predict neural activity and behavioral responses for images it has not been trained on.

Box 2 | Neural network models

The term "neural network model" has come to be associated with a class of model that is inspired by biological neural networks in that each unit combines many inputs and information is processed in parallel through a network. In contrast to biologically detailed models, which may capture action potentials and dynamics in multiple compartments of each neuron, these models abstract from the biological details. However, they can explain certain cognitive functions, such as visual object recognition, and therefore provide an attractive framework for linking cognition to the brain.

A typical unit computes a linear combination of its inputs and passes the result through a static nonlinearity. The output is sometimes interpreted as analogous to the firing rate of a neuron. Even *shallow* networks (those with a single layer of hidden units between inputs and outputs) can approximate arbitrary functions[123]. However, *deep* networks (those with multiple hidden layers) can more efficiently capture many of the complex functions needed in real-world tasks. Many applications—for example, in computer vision—use feedforward architectures. However, recurrent neural networks, which reprocess the outputs of their units and generate complex dynamics, have brought additional engineering advances[76] and better capture the recurrent signaling in brains[35,124–126]. Whereas feedforward networks are universal function approximators, recurrent networks are universal approximators of dynamical systems[127]. Recurrent processing enables a network to recycle its limited computational resources through time so as to perform more complex sequences of computations. Recurrent networks can represent the recent stimulus history in a dynamically compressed format, providing the temporal context information needed for current processing. As a result, recurrent networks can recognize, predict, and generate dynamical patterns.

Both feedforward and recurrent networks are defined by their architecture and the setting of the connection weights. One way to set the weights is through iterative small adjustments that bring the output closer to some desired output (supervised learning). Each weight is adjusted in proportion to the reduction in the error that a small change to it would yield. This method is called *gradient descent* because it produces steps in the space of weights along which the error declines most steeply. Gradient descent can be implemented using *backpropagation*, an efficient algorithm for computing the derivative of the error function with respect to each weight.

Whether the brain uses an algorithm like backpropagation for learning is controversial. Several biologically plausible implementations of backpropagation or closely related forms of supervised learning have been suggested[128–130]. Supervision signals might be generated internally[131] on the basis of the context provided by multiple sensory modalities; on the basis of the dynamic refinement of representations over time, as more evidence becomes available from the senses and from memory[132]; and on the basis of internal and external reinforcement signals arising in interaction with the environment[133]. Reinforcement learning[41] and unsupervised learning of neural network parameters[119,134] are areas of rapid current progress.

Neural network models have demonstrated that taking inspiration from biology can yield breakthroughs in AI. It seems likely that the quest for models that can match human cognitive abilities will draw us deeper into the biology[135]. The abstract neural network models currently most successful in engineering could be implemented with biological hardware. However, they only use a small subset of the dynamical components of brains. Neuroscience has described a rich repertoire of dynamical components, including action potentials[108], canonical microcircuits[136], dendritic dynamics[128,130,137] and network phenomena[27], such as oscillations[138], that may have computational functions. Biology also provides constraints on the global architecture, suggesting, for example, complementary subsystems for learning[139]. Modeling these biological components in the context of neural networks designed to perform meaningful tasks may reveal how they contribute to brain computation and may drive further advances in AI.

Several recent studies have begun to test neural network models as models of brain information processing[37,38]. These studies predicted brain representations of novel images in the primate ventral visual stream with deep convolutional neural network models trained to recognize objects in images. Results have shown that the internal representations of deep convolutional neural networks provide the best current models of representations of visual images in inferior temporal cortex in humans and monkeys[77–79]. When comparing large numbers of models, those that were optimized to perform the task of object classification better explained the cortical representation[77,78].

Early layers of deep neural networks trained to recognize objects contain representations resembling those in early visual cortex[78,80]. As we move along the ventral visual stream, higher layers of the neural networks come to provide a better basis for explaining the representations[80–82]. Higher layers of deep convolutional neural networks also resemble the inferior temporal cortical representation in that both enable the decoding of object position, size and pose, along with the category of the object[83]. In addition to testing these models by predicting brain-activity data, the field has begun to test them by predicting behavioral responses reflecting perceived shape[84] and object similarity[85].

Cognitive models. Models at the cognitive level enable researchers to envision the information processing without simultaneously having to tackle its implementation with neurobiologically plausible components. This enables progress on domains of higher cognition, where neural network models still fall short. Moreover, a cognitive model may provide a useful abstraction, even when a process can also be captured with a neural network model.

Neuroscientific explanations now dominate for functional components closer to the periphery of the brain, where sensory and motor processes connect the animal to its environment. However, much of higher-level cognition has remained beyond the reach of neuroscientific accounts and neural network models. To illustrate some of the unique contributions of cognitive models, we briefly discuss three classes of cognitive model: production systems, reinforcement learning models and Bayesian cognitive models.

Production systems provide an early example of a class of cognitive models that can explain reasoning and problem solving. These models use rules and logic, and are symbolic in that they operate on symbols rather than sensory data and motor signals. They capture cognition, rather than perception and motor control, which ground cognition in the physical environment. A 'production' is a cognitive action triggered according to an if-then rule. A set of such rules specifies the conditions ('if') under which each of a range of productions ('then') is to be executed. The conditions refer to current goals and knowledge in memory. The actions can modify the internal state of goals and knowledge. For example, a production may create a subgoal or store an inference. If conditions are met for multiple

Box 3 | Bayesian cognitive models

Bayesian cognitive models are motivated by the assumption that the brain approximates the statistically optimal solution to a task. The statistically optimal way to make inferences and decide what to do is to interpret the current sensory evidence in light of all available prior knowledge using the rules of probability. Consider the case of visual perception. The retinal signals reflect the objects in the world, which we would like to recognize. To infer the objects, we should consider what configurations of objects we deem possible and how well each explains the image. Our prior beliefs are represented by a *generative model* that captures the probability of each configuration of objects and the probabilities with which a given configuration would produce different retinal images.

More formally, a Bayesian model of vision might use a generative model of the joint distribution $p(\mathbf{d}, \mathbf{c})$ of the sensory data $\mathbf{d}$ (the image) and the causes in the world $\mathbf{c}$ (the configuration of surfaces, objects and light sources to be inferred)[140]. The joint distribution $p(\mathbf{d}, \mathbf{c})$ equals the product of the *prior*, $p(\mathbf{c})$, over all possible configurations of causes and the *likelihood*, $p(\mathbf{d}|\mathbf{c})$, the probability of a particular image given a particular configuration of causes. A prescribed model for $p(\mathbf{d}|\mathbf{c})$ would enable us to evaluate the likelihood, the probability of a specific image $\mathbf{d}$ given specific causes $\mathbf{c}$. Alternatively, we might have an implicit model for $p(\mathbf{d}|\mathbf{c})$ in the form of a stochastic mapping from causes $\mathbf{c}$ to data $\mathbf{d}$ (images). Such a model would generate natural images. Whether prescribed or implicit, the model of $p(\mathbf{d}|\mathbf{c})$ captures how the causes in the world create the image, or at least how they relate to the image. Visual recognition amounts to computing the *posterior* $p(\mathbf{c}|\mathbf{d})$, the probability distribution over the causes given a particular image. The posterior $p(\mathbf{c}|\mathbf{d})$ reveals the causes $\mathbf{c}$ as they would have to exist in the world to explain the sensory data $\mathbf{d}$[141]. A model computing $p(\mathbf{c}|\mathbf{d})$ is called a *discriminative model* because it discriminates among images—here mapping from effects (the image) to the causes. The inversion mathematically requires a prior $p(\mathbf{c})$ over the latent causes. The prior $p(\mathbf{c})$ can constrain the interpretation and help reduce the ambiguity resulting from the multiple configurations of causes that can account for any image.

Basing the inference of the causes $\mathbf{c}$ on a generative model of $p(\mathbf{d}, \mathbf{c})$ that captures all available knowledge and uncertainty is statistically optimal (i.e., it provides the best inferences given limited data), but computationally challenging (i.e., it may require more neurons or time than the animal can use). Ideally, the generative model $p(\mathbf{d}, \mathbf{c})$ implicit to the inference $p(\mathbf{c}|\mathbf{d})$ should capture our knowledge not just about image formation, but also the things in the world and their interactions, and our uncertainties about these processes. One challenge is to learn a generative model from sensory data. We need to represent the learned knowledge and the remaining uncertainties. If the generative model is mis-specified, then the inference will not be optimal. For real-world tasks, some degree of misspecification of the model is inevitable. For example, the generative model may contain an overly simplified version of the image-generation process. Another challenge is the computation of the posterior $p(\mathbf{c}|\mathbf{d})$. For realistically complex generative models, the inference may require computationally intensive iterative algorithms such as *Markov chain Monte Carlo*, *belief propagation* or *variational inference*. The brain's compromise between statistical and computational efficiency[142–144] may involve learning fast feedforward recognition models that speed up frequent component inferences, crystallizing conclusions that are costly to fluidly derive with iterative algorithms. This is known as amortized inference[145,146].

Bayesian cognitive models have recently flourished in interaction with machine learning and statistics. Early work used generative models with a fixed structure that were flexible only with respect to a limited set of parameters. Modern generative models can grow in complexity with the data and discover their inherent structure[98]. They are called *nonparametric* because they are not limited by a predefined finite set of parameters[147]. Their parameters can grow in number without any predefined bound.

rules, a conflict-resolution mechanism chooses one production. A model specified using this formalism will generate a sequence of productions, which may to some extent resemble our conscious stream of thought while working toward some cognitive goal. The formalism of production systems also provides a universal computational architecture[86]. Production systems such as ACT-R[5] were originally developed under the guidance of behavioral data. More recently such models have also begun to be tested in terms of their ability to predict regional-mean fMRI activation time courses[87].

Reinforcement learning models capture how an agent can learn to maximize its long-term cumulative reward through interaction with its environment[88,89]. As in production systems, reinforcement learning models often assume that the agent has perception and motor modules that enable the use of discrete symbolic representations of states and actions. The agent chooses actions, observes resulting states of the environment, receives rewards along the way and learns to improve its behavior. The agent may learn a 'value function' associating each state with its expected cumulative reward. If the agent can predict which state each action leads to and if it knows the values of those states, then it can choose the most promising action. The agent may also learn a 'policy' that associates each state directly with promising actions. The choice of action must balance exploitation (which brings short-term reward) and exploration (which benefits learning and brings long-term reward).

The field of reinforcement learning explores algorithms that define how to act and learn so as to maximize cumulative reward. With roots in psychology and neuroscience, reinforcement learning theory is now an important field of machine learning and AI. It provides a very general perspective on control that includes the classical techniques dynamic programming, Monte Carlo and exhaustive search as limiting cases, and can handle challenging scenarios in which the environment is stochastic and only partially observed, and its causal mechanisms are unknown.

An agent might exhaustively explore an environment and learn the most promising action to take in any state by trial and error (model-free control). This would require sufficient time to learn, enough memory, and an environment that does not kill the agent prematurely. Biological organisms, however, have limited time to learn and limited memory, and must avoid interactions that might kill them. Under these conditions, an agent might do better to build a model of its environment. A model can compress and generalize experience to enable intelligent action in novel situations (model-based control). Model-free methods are computationally efficient (mapping from states to values or directly to actions), but statistically inefficient (learning takes long); model-based methods are more statistically efficient, but may require prohibitive amounts of computation (to simulate possible futures)[90].

Until experience is sufficient to build a reliable model, an agent might do best to simply store episodes and revert to paths of action that have met with success in the past (episodic control)[91,92]. Storing episodes preserves sequential dependency information important for model building. Moreover, episodic

Box 4 | Why do cognitive science, computational neuroscience and AI need one another?

Cognitive science needs computational neuroscience, not merely to explain the implementation of cognitive models in the brain, but also to discover the algorithms. For example, the dominant models of sensory processing and object recognition are brain-inspired neural networks, whose computations are not easily captured at a cognitive level. Recent successes with Bayesian nonparametric models do not yet in general scale to real-world cognition. Explaining the computational efficiency of human cognition and predicting detailed cognitive dynamics and behavior could benefit from studying brain-activity dynamics. Explaining behavior is essential, but behavioral data alone provide insufficient constraints for complex models. Brain data can provide rich constraints for cognitive algorithms if leveraged appropriately. Cognitive science has always progressed in close interaction with artificial intelligence. The disciplines share the goal of building task-performing models and thus rely on common mathematical theory and programming environments.

Computational neuroscience needs cognitive science to challenge it to engage higher-level cognition. At the experimental level, the tasks of cognitive science enable computational neuroscience to bring cognition into the lab. At the level of theory, cognitive science challenges computational neuroscience to explain how the neurobiological dynamical components it studies contribute to cognition and behavior. Computational neuroscience needs AI, and in particular machine learning, to provide the theoretical and technological basis for modeling cognitive functions with biologically plausible dynamical components.

Artificial intelligence needs cognitive science to guide the engineering of intelligence. Cognitive science's tasks can serve as benchmarks for AI systems, building up from elementary cognitive abilities to artificial general intelligence. The literatures on human development and learning provide an essential guide to what is possible for a learner to achieve and what kinds of interaction with the world can support the acquisition of intelligence. AI needs computational neuroscience for algorithmic inspiration. Neural network models are an example of a brain-inspired technology that is unrivalled in several domains of AI. Taking further inspiration from the neurobiological dynamical components (for example, spiking neurons, dendritic dynamics, the canonical cortical microcircuit, oscillations, neuromodulatory processes) and the global functional layout of the human brain (for example, subsystems specialized for distinct functions, including sensory modalities, memory, planning and motor control) might lead to further AI breakthroughs. Machine learning draws from separate traditions in statistics and computer science, which have optimized statistical and computational efficiency, respectively. The integration of computational and statistical efficiency is an essential challenge in the age of big data. The brain appears to combine computational and statistical efficiency, and understanding its algorithm might boost machine learning.

control enables the agent to exploit such dependencies even before understanding the causal mechanism supporting a successful path of action.

The brain is capable of each of these three modes of control (model-free, model-based, episodic)[89] and appears to combine their advantages using an algorithm that has yet to be discovered. AI and computational neuroscience share the goal of discovering this algorithm[41,90,93–95], although they approach this goal from different angles. This is an example of how a cognitive challenge can motivate the development of formal models and drive progress in AI and neuroscience.

A third, and critically important, class of cognitive model is that of Bayesian models (Box 3)[21,96–98]. Bayesian inference provides an essential normative perspective on cognition. It tells us what a brain should in fact compute for an animal to behave optimally. Perceptual inference, for example, should consider the current sensory data in the context of prior beliefs. Bayesian inference simply refers to combining the data with prior beliefs according to the rules of probability.

Bayesian models have contributed to our understanding of basic sensory and motor processes[22–24]. They have also provided insights into higher cognitive processes of judgment and decision making, explaining classical cognitive biases[99] as the product of prior assumptions, which may be incorrect in the experimental task but correct and helpful in the real world.

With Bayesian nonparametric models, cognitive science has begun to explain more complex cognitive abilities. Consider the human ability to induce a new object category from a single example. Such inductive inference requires prior knowledge of a kind not captured by current feedforward neural network models[100]. To induce a category, we rely on an understanding of the object, of the interactions among its parts, of how they give rise to its function. In the Bayesian cognitive perspective, the human mind, from infancy, builds mental models of the world[2]. These models may not only be generative models in the probabilistic sense, but may be causal and compositional, supporting mental simulations of processes in the world using elements that can be re-composed to generalize to novel and hypothetical scenarios[2,98,101]. This modeling approach has been applied to our reasoning about the physical[101–103] and even the social[104] world.

Generative models are an essential ingredient of general intelligence. An agent attempting to learn a generative model strives to understand all relationships among its experiences. It does not require external supervision or reinforcement to learn, but can mine all its experiences for insights on its environment and itself. In particular, causal models of processes in the world (how objects cause images, how the present causes the future) can give an agent a deeper understanding and thus a better basis for inferences and actions.

The representation of probability distributions in neuronal populations has been explored theoretically and experimentally[105,106]. However, relating Bayesian inference and learning, especially structure learning in nonparametric models, to its implementation in the brain remains challenging[107]. As theories of brain computation, approximate inference algorithms such as sampling may explain cortical feedback signals and activity correlations[97,108–110]. Moreover, the corners cut by the brain for computational efficiency, the approximations, may explain human deviations from statistical optimality. In particular, cognitive experiments have revealed signatures of sampling[111] and amortized inference[112] in human behavior.

Cognitive models, including the three classes highlighted here, decompose cognition into meaningful functional components. By declaring their models independent of the implementation in the brain, cognitive scientists are able to address high-level cognitive processes[21,97,98] that are beyond the reach of current neural networks. Cognitive models are essential for cognitive computational neuroscience because they enable us to see the whole as we attempt to understand the roles of the parts.

Box 5 | Shareable tasks, data, models and tests: a new culture of multidisciplinary collaboration

Neurobiologically plausible models that explain cognition will have substantial parametric complexity. Building and evaluating such models will require machine learning and big brain and behavioral datasets. Traditionally, each lab has developed its own tasks, datasets, models and tests with a focus on the goals of its own discipline. To scale these efforts up to meet the challenge, we will need to develop tasks, data, models and tests that are relevant across the three disciplines and shared among labs (see figure). A new culture of collaboration will assemble big data and big models by combining components from different labs. To meet the conjoined criteria for success of cognitive science, computational neuroscience and artificial intelligence, the best division of labor might cut across the traditional disciplines.

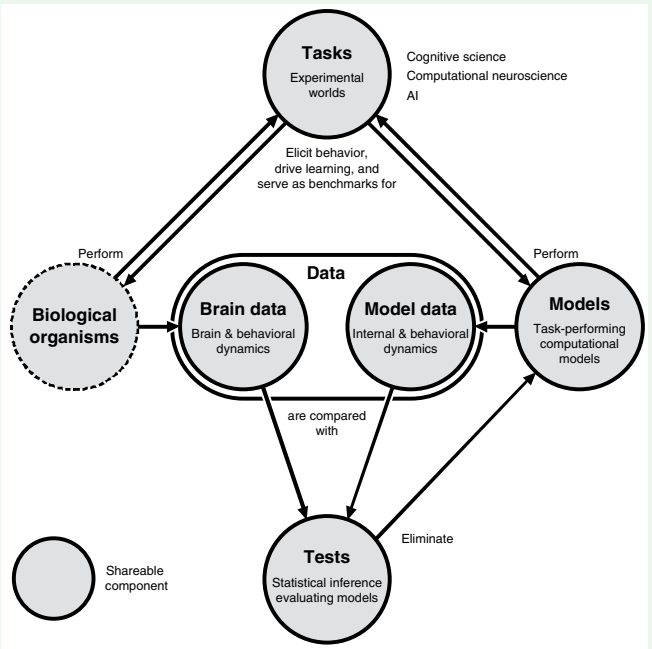
Tasks. By designing experimental tasks, we carve up cognition into components that can be quantitatively investigated. A task is a controlled environment for behavior. It defines the dynamics of a task 'world' that provides sensory input (for example, visual stimuli) and captures motor output (for example, button press, joystick control or higher-dimensional limb or whole-body control). Tasks drive the acquisition of brain and behavioral data and the development of AI models, providing well-defined challenges and quantitative performance benchmarks for comparing models. The ImageNet tasks[148], for example, have driven substantial progress in computer vision. Tasks should be designed and implemented such that they can readily be used in all three disciplines to drive data acquisition and model development (related developments include OpenAI's Gym, https://gym.openai.com/; Universe, https://universe.openai.com/; and DeepMind's Lab[149]). The spectrum of useful tasks includes classical psychophysical tasks employing simple stimuli and responses as well as interactions in virtual realities. As we engage all aspects of the human mind, our tasks will need to simulate natural environments and will come to resemble computer games. This may bring the added benefit of mass participation and big behavioral data, especially when tasks are performed via the Internet[150].

Data. Behavioral data acquired during task performance provides overall performance estimates and detailed signatures of success and failure, of reaction times and movement trajectories. Brain-activity measurements characterize the dynamic computations underlying task performance. Anatomical data can characterize the structure and connectivity of the brain at multiple scales. Structural brain data, functional brain data and behavioral data will all be essential for constraining computational models.

Models. Task-performing computational models can take sensory inputs and produce motor outputs so as to perform experimental tasks. AI-scale neurobiologically plausible models can be shared openly and tested in terms of their task performance and in terms of their ability to explain a variety of brain and behavioral datasets, including new datasets acquired after definition of the model. Initially, many models will be specific to small subsets of tasks. Ultimately, models must generalize across tasks.

Tests. To assess the extent to which a model can explain brain information processing during a particular task, we need tests that compare models and brains on the basis of brain and behavioral data. Every brain is idiosyncratic in its structure and function. Moreover, for a given brain, every act of perception, cognition and action is unique in time and cannot be repeated precisely because it permanently changes the brain in question. These complications make it challenging to compare brains and models. We must define the summary statistics of interest and the correspondence mapping between model and brain in space and time at some level of abstraction. Developing appropriate tests for adjudicating among models and determining how close we are to understanding the brain is not merely a technical challenge of statistical inference. It is a conceptual challenge fundamental to theoretical neuroscience.

The interaction among labs and disciplines can benefit from adversarial cooperation[134]. Cognitive researchers who feel that current computational models fall short of explaining an important aspect of cognition are challenged to design shareable tasks and tests that quantify these shortcomings and to provide human behavioral data to set the bar for AI models. Neuroscientists who feel that current models do not explain brain information processing are challenged to share brain-activity data acquired during task performance and tests comparing activity patterns between brains and models to quantify the shortcomings of the models. Although we will have a plurality of definitions of success, translating these into quantitative measures of the quality of a model is essential and could drive progress in cognitive computational neuroscience, as well as engineering.



Interactions among shareable components. Tasks, data, models and tests are components (gray nodes) that lend themselves to sharing among labs and across disciplines, to enable collaborative construction and testing of big models driven by big brain and behavioral datasets assembled across labs.

Looking ahead

Bottom up and top down. The brain seamlessly merges bottom-up discriminative and top-down generative computations in perceptual inference, and model-free and model-based control. Brain science likewise needs to integrate its levels of description and to progress both bottom-up and top-down, so as to explain task

performance on the basis of neuronal dynamics and provide a mechanistic account of how the brain gives rise to the mind.

Bottom-up visions, proceeding from detailed measurements toward an understanding of brain computation, have been prominent and have driven the most important recent funding initiatives. The European Human Brain Project and the US BRAIN Initiative are both motivated by bottom-up visions, in which an understanding of brain computation is achieved by measuring and modeling brain dynamics with a focus on the circuit level. The BRAIN Initiative seeks to advance technologies for measuring and manipulating neuronal activity. The Human Brain Project attempts to synthesize neuroscience data in biologically detailed dynamic models. Both initiatives proceed primarily from experiment toward theory and from the cellular level of description to larger-scale phenomena.

Measuring large numbers of neurons simultaneously and modeling their interactions at the circuit level will be essential. The bottom-up vision is grounded in the history of science. Microscopes and telescopes, for example, have brought scientific breakthroughs. However, it is always in the context of prior theory (generative models of the observed processes) that better observations advance our understanding. In astronomy, for example, the theory of Copernicus guided Galileo in interpreting his telescopic observations.

Understanding the brain requires that we develop theory and experiment in tandem and complement the bottom-up, data-driven approach by a top-down, theory-driven approach that starts with behavioral functions to be explained[113,114]. Unprecedentedly rich measurements and manipulations of brain activity will drive theoretical insight when they are used to adjudicate between brain-computational models that pass the first test of being able to perform a function that contributes to the behavioral fitness of the organism. The top-down approach, therefore, is an essential complement to the bottom-up approach toward understanding the brain (Fig. 3).

Integrating Marr's levels. Marr (1982) offered a distinction of three levels of analysis: (i) computational theory, (ii) representation and algorithm, and (iii) neurobiological implementation[115]. Cognitive science starts from computational theory, decomposing cognition into components and developing representations and algorithms from the top down. Computational neuroscience proceeds from the bottom up, composing neuronal building blocks into representations and algorithms thought to be useful components in the context of the brain's overall function. AI builds representations and algorithms that combine simple components to implement complex feats of intelligence. All three disciplines thus converge on the algorithms and representations of the brain and mind, contributing complementary constraints[116].

Marr's levels provide a useful guide to the challenge of understanding the brain. However, they should not be taken to suggest that cognitive science need not consider the brain or that computational neuroscience need not consider cognition (Box 4). Marr was inspired by computers, which are designed by human engineers to precisely conform to high-level algorithmic descriptions. This enables the engineers to abstract from the circuits when designing the algorithms. Even in computer science, however, certain aspects of the algorithms depend on the hardware, such as its parallel processing capabilities. Brains differ from computers in ways that exacerbate this dependence. Brains are the product of evolution and development, processes that are not constrained to generate systems whose behavior can be perfectly captured at some abstract level of description. It may therefore not be possible to understand cognition without considering its implementation in the brain or, conversely, to make sense of neuronal circuits except in the context of the cognitive functions they support.

For an example of a challenge that transcends the disciplines, consider a child seeing an escalator for the first time. She will rapidly recognize people on steps traveling upward obliquely. She might think of it as a moving staircase and imagine riding on it, being lifted one story without exerting any effort. She might infer its function and form a new concept on the basis of a single experience, before ever learning the word "escalator".

Deep neural network models provide a biologically plausible account of the rapid recognition of the elements of the visual experience (people, steps, oblique upward motion, handrail). They can explain the computationally efficient pattern recognition component[42]. However, they cannot explain yet how the child understands the relationships among the elements, the physical interactions of the objects, the people's goal to go up, and the function of the escalator, or how she can imagine the experience and instantly form a new concept.

Bayesian nonparametric models explain how deep inferences and concept formation from single experiences are even possible. They may explain the brain's stunning statistical efficiency, its ability to infer so much from so little data by building generative models that provide abstract prior knowledge[98]. However, current inference algorithms require large amounts of computation and, as a result, do not yet scale to real-world challenges such as forming the new concept "escalator" from a single visual experience.

On a 20-watt power budget, the brain's algorithms combine statistical and computational efficiency in ways that are beyond current AI of either the Bayesian or the neural network variety. However, recent work in AI and machine learning has begun to explore the intersection between Bayesian inference and neural network models, combining the statistical strengths of the former (uncertainty representation, probabilistic inference, statistical efficiency) with the computational strengths of the latter (representational learning, universal function approximation, computational efficiency)[117–119].

Integrating all three of Marr's levels will require close collaboration among researchers with a wide variety of expertise. It is difficult for any single lab to excel at neuroscience, cognitive science and AI-scale computational modeling. We therefore need collaborations between labs with complementary expertise. In addition to conventional collaborations, an open science culture, in which components are shared between disciplines, can help us integrate Marr's levels. Shareable components include cognitive tasks, brain and behavioral data, computational models, and tests that evaluate models by comparing them to biological systems (Box 5).

The study of the mind and brain is entering a particularly exciting phase. Recent advances in computer hardware and software enable AI-scale modeling of the mind and brain. If cognitive science, computational neuroscience and AI can come together, we might be able to explain human cognition with neurobiologically plausible computational models.

Received: 7 November 2016; Accepted: 11 July 2018;
Published online: 20 August 2018

References

1. Newell, A. You can't play 20 questions with nature and win: projective comments on the papers of this symposium. Technical Report, School of Computer Science, Carnegie Mellon University (1973).
2. 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).
3. Kriegeskorte, N. & Mok, R. M. Building machines that adapt and compute like brains. *Behav. Brain Sci.* **40**, e269 (2017).
4. Simon, H. A. & Newell, A. Human problem solving: the state of the theory in 1970. *Am. Psychol.* **26**, 145–159 (1971).
5. Anderson, J. R. *The Architecture of Cognition* (Harvard Univ. Press, Cambridge, MA, USA, 1983).

6. McClelland, J. L. & Rumelhart, D. E. *Parallel Distributed Processing* (MIT Press, Cambridge, MA, USA, 1987).
7. Gazzaniga, M. S. ed. *The Cognitive Neurosciences* (MIT Press, Cambridge, MA, USA, 2004).
8. Fodor, J. A. Précis of The Modularity of Mind. *Behav. Brain Sci.* **8**, 1 (1985).
9. Chklovskii, D. B. & Koulakov, A. A. Maps in the brain: what can we learn from them? *Annu. Rev. Neurosci.* **27**, 369–392 (2004).
10. Szucs, D. & Ioannidis, J. P. A. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol.* **15**, e2000797 (2017).
11. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. & Baker, C. I. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* **12**, 535–540 (2009).
12. Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
13. Tsao, D. Y., Freiwald, W. A., Tootell, R. B. & Livingstone, M. S. A cortical region consisting entirely of face-selective cells. *Science* **311**, 670–674 (2006).
14. Freiwald, W. A. & Tsao, D. Y. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* **330**, 845–851 (2010).
15. Grill-Spector, K., Weiner, K. S., Kay, K. & Gomez, J. The functional neuroanatomy of human face perception. *Annu. Rev. Vis. Sci.* **3**, 167–196 (2017).
16. Yildirim, I. et al. Efficient and robust analysis-by-synthesis in vision: a computational framework, behavioral tests, and modeling neuronal representations. in *Annual Conference of the Cognitive Science Society* (eds. Noelle, D. C. et al.) (Cognitive Science Society, Austin, TX, USA, 2015).
17. Kriegeskorte, N., Formisano, E., Sorger, B. & Goebel, R. Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proc. Natl Acad. Sci. USA* **104**, 20600–20605 (2007).
18. Anzellotti, S., Fairhall, S. L. & Caramazza, A. Decoding representations of face identity that are tolerant to rotation. *Cereb. Cortex* **24**, 1988–1995 (2014).
19. Chang, L. & Tsao, D. Y. The code for facial identity in the primate brain. *Cell* **169**, 1013–1028.e14 (2017).
20. Van Essen, D. C. et al. The Brain Analysis Library of Spatial maps and Atlases (BALSA) database. *Neuroimage* **144**(Pt. B), 270–274 (2017).
21. Griffiths, T. L., Chater, N., Kemp, C., Perfors, A. & Tenenbaum, J. B. Probabilistic models of cognition: exploring representations and inductive biases. *Trends Cogn. Sci.* **14**, 357–364 (2010).
22. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
23. Weiss, Y., Simoncelli, E. P. & Adelson, E. H. Motion illusions as optimal percepts. *Nat. Neurosci.* **5**, 598–604 (2002).
24. Körding, K. P. & Wolpert, D. M. Bayesian integration in sensorimotor learning. *Nature* **427**, 244–247 (2004).
25. MacKay, D. J. C. *Information Theory, Inference, and Learning Algorithms*. (Cambridge Univ. Press, Cambridge, 2003)
26. Murphy, K. P. *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, MA, USA, 2012).
27. Dayan, P. & Abbott, L. F. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (MIT Press, Cambridge, MA, USA, 2001).
28. Abbott, L. F. Theoretical neuroscience rising. *Neuron* **60**, 489–495 (2008).
29. Olshausen, B. A. & Field, D. J. Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* **14**, 481–487 (2004).
30. Simoncelli, E. P. & Olshausen, B. A. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).
31. Carandini, M. & Heeger, D. J. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2011).
32. Chaudhuri, R. & Fiete, I. Computational principles of memory. *Nat. Neurosci.* **19**, 394–403 (2016).
33. Shadlen, M. N. & Kiani, R. Decision making as a window on cognition. *Neuron* **80**, 791–806 (2013).
34. Newsome, W. T., Britten, K. H. & Movshon, J. A. Neuronal correlates of a perceptual decision. *Nature* **341**, 52–54 (1989).
35. Wang, X.-J. Decision making in recurrent neuronal circuits. *Neuron* **60**, 215–234 (2008).
36. Diedrichsen, J., Shadmehr, R. & Ivry, R. B. The coordination of movement: optimal feedback control and beyond. *Trends Cogn. Sci.* **14**, 31–39 (2010).
37. Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015).
38. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
39. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. in *Advances in Neural Information Processing Systems 25* 1097–1105 (Curran Associates, Red Hook, NY, USA, 2012).
40. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
41. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
42. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
43. Cohen, J. D. et al. Computational approaches to fMRI analysis. *Nat. Neurosci.* **20**, 304–313 (2017).
44. Forstmann, B. U., Wagenmakers, E.-J., Eichele, T., Brown, S. & Serences, J. T. Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? *Trends Cogn. Sci.* **15**, 272–279 (2011).
45. Deco, G., Tononi, G., Boly, M. & Kringelbach, M. L. Rethinking segregation and integration: contributions of whole-brain modelling. *Nat. Rev. Neurosci.* **16**, 430–439 (2015).
46. Biswal, B., Yetkin, F. Z., Haughton, V. M. & Hyde, J. S. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.* **34**, 537–541 (1995).
47. Hyvarinen, A., Karhunen, J. & Oja, E. *Independent Component Analysis* (Wiley, Hoboken, NJ, USA, 2001).
48. Bullmore, E. T. & Bassett, D. S. Brain graphs: graphical models of the human brain connectome. *Annu. Rev. Clin. Psychol.* **7**, 113–140 (2011).
49. Deco, G., Jirsa, V. K. & McIntosh, A. R. Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nat. Rev. Neurosci.* **12**, 43–56 (2011).
50. Friston, K. Dynamic causal modeling and Granger causality. Comments on: the identification of interacting networks in the brain using fMRI: model selection, causality and deconvolution. *Neuroimage* **58**, 303–305 (2011). author reply 310–311.
51. Dennett, D. C. *The Intentional Stance* (MIT Press, Cambridge, MA, USA, 1987).
52. Diedrichsen, J. & Kriegeskorte, N. Representational models: a common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Comput. Biol.* **13**, e1005508 (2017).
53. Afraz, S.-R., Kiani, R. & Esteky, H. Microstimulation of inferotemporal cortex influences face categorization. *Nature* **442**, 692–695 (2006).
54. Parvizi, J. et al. Electrical stimulation of human fusiform face-selective regions distorts face perception. *J. Neurosci.* **32**, 14915–14920 (2012).
55. Norman, K. A., Polyn, S. M., Detre, G. J. & Haxby, J. V. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* **10**, 424–430 (2006).
56. Tong, F. & Pratte, M. S. Decoding patterns of human brain activity. *Annu. Rev. Psychol.* **63**, 483–509 (2012).
57. Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* **17**, 401–412 (2013).
58. Haxby, J. V., Connolly, A. C. & Guntupalli, J. S. Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci.* **37**, 435–456 (2014).
59. Haynes, J.-D. A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. *Neuron* **87**, 257–270 (2015).
60. Jin, X. & Costa, R. M. Shaping action sequences in basal ganglia circuits. *Curr. Opin. Neurobiol.* **33**, 188–196 (2015).
61. DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. *Trends Cogn. Sci.* **11**, 333–341 (2007).
62. Naselaris, T. & Kay, K. N. Resolving ambiguities of MVPA using explicit models of representation. *Trends Cogn. Sci.* **19**, 551–554 (2015).
63. Mitchell, T. M. et al. Predicting human brain activity associated with the meanings of nouns. *Science* **320**, 1191–1195 (2008).
64. Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).
65. Dumoulin, S. O. & Wandell, B. A. Population receptive field estimates in human visual cortex. *Neuroimage* **39**, 647–660 (2008).
66. Diedrichsen, J., Ridgway, G. R., Friston, K. J. & Wiestler, T. Comparing the similarity and spatial structure of neural representations: a pattern-component model. *Neuroimage* **55**, 1665–1678 (2011).
67. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
68. Nili, H. et al. A toolbox for representational similarity analysis. *PLoS Comput. Biol.* **10**, e1003553 (2014).
69. Devereux, B. J., Clarke, A., Marouchos, A. & Tyler, L. K. Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *J. Neurosci.* **33**, 18906–18916 (2013).

70. Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
71. Markram, H. The Blue Brain Project. *Nat. Rev. Neurosci.* **7**, 153–160 (2006).
72. Eliasmith, C. & Trujillo, O. The use and abuse of large-scale brain models. *Curr. Opin. Neurobiol.* **25**, 1–6 (2014).
73. Eliasmith, C. et al. A large-scale model of the functioning brain. *Science* **338**, 1202–1205 (2012).
74. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron* **95**, 245–258 (2017).
75. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
76. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, Cambridge, MA, USA, 2016).
77. 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).
78. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
79. Cadieu, C. F. et al. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* **10**, e1003963 (2014).
80. Güçlü, 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).
81. Eickenberg, M., Gramfort, A., Varoquaux, G. & Thirion, B. Seeing it all: convolutional network layers map the function of the human visual system. *Neuroimage* **152**, 184–194 (2017).
82. 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).
83. Hong, H., Yamins, D. L. K., Majaj, N. J. & DiCarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* **19**, 613–622 (2016).
84. Kubilius, J., Bracci, S. & Op de Beeck, H. P. Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol.* **12**, e1004896 (2016).
85. Jozwik, K. M., Kriegeskorte, N., Storrs, K. R. & Mur, M. Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Front. Psychol.* **8**, 1726 (2017).
86. Moore, C. & Mertens, S. *The Nature of Computation*. (Oxford Univ. Press, Oxford, 2011).
87. Borst, J., Taatgen & Anderson, J. Using the ACT-R cognitive architecture in combination with fMRI data. in *An Introduction to Model-Based Cognitive Neuroscience* (eds. Forstmann, B. U. & Wagenmakers, E.-J.) (Springer, New York, 2014).
88. Sutton, R. & Barto, A. *Reinforcement Learning: An Introduction* Vol. 1 (MIT Press, Cambridge, MA, USA, 1998).
89. O'Doherty, J. P., Cockburn, J. & Pauli, W. M. Learning, reward, and decision making. *Annu. Rev. Psychol.* **68**, 73–100 (2017).
90. Daw, N. D. & Dayan, P. The algorithmic anatomy of model-based evaluation. *Phil. Trans. R. Soc. Lond. B* **369**, 20130478 (2014).
91. Lengyel, M. & Dayan, P. Hippocampal contributions to control: the third way in *Advances in Neural Information Processing Systems 20* 889–896 (MIT Press, Cambridge, MA, USA, 2008)..
92. Gershman, S. J. & Daw, N. D. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annu. Rev. Psychol.* **68**, 101–128 (2017).
93. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
94. Sutton, R. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. in *Proceedings of the Seventh International Conference on Machine Learning* 216–224 (Morgan Kaufmann, San Francisco, 1990).
95. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
96. Ma, W. J. Organizing probabilistic models of perception. *Trends Cogn. Sci.* **16**, 511–518 (2012).
97. Fiser, J., Berkes, P., Orbán, G. & Lengyel, M. Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn. Sci.* **14**, 119–130 (2010).
98. 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).
99. Tversky, A. & Kahneman, D. Judgment under uncertainty: heuristics and biases. in *Utility, Probability, and Human Decision Making* (eds. Wendt, D. & Vlek, C.) 141–162, https://doi.org/10.1007/978-94-010-1834-0_8 (Springer Netherlands, Dordrecht, the Netherlands, 1975).
100. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
101. Ullman, T. D., Spelke, E., Battaglia, P. & Tenenbaum, J. B. Mind games: game engines as an architecture for intuitive physics. *Trends Cogn. Sci.* **21**, 649–665 (2017).
102. Battaglia, P. W., Hamrick, J. B. & Tenenbaum, J. B. Simulation as an engine of physical scene understanding. *Proc. Natl Acad. Sci. USA* **110**, 18327–18332 (2013).
103. Kubricht, J. R., Holyoak, K. J. & Lu, H. Intuitive physics: current research and controversies. *Trends Cogn. Sci.* **21**, 749–759 (2017).
104. Pantelis, P. C. et al. Inferring the intentional states of autonomous virtual agents. *Cognition* **130**, 360–379 (2014).
105. Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* **16**, 1170–1178 (2013).
106. Orhan, A. E. & Ma, W. J. Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback. *Nat. Commun.* **8**, 138 (2017).
107. Tervo, D. G. R., Tenenbaum, J. B. & Gershman, S. J. Toward the neural implementation of structure learning. *Curr. Opin. Neurobiol.* **37**, 99–105 (2016).
108. Buesing, L., Bill, J., Nessler, B. & Maass, W. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Comput. Biol.* **7**, e1002211 (2011).
109. Haefner, R. M., Berkes, P. & Fiser, J. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron* **90**, 649–660 (2016).
110. Aitchison, L. & Lengyel, M. The Hamiltonian brain: efficient probabilistic inference with excitatory-inhibitory neural circuit dynamics. *PLoS Comput. Biol.* **12**, e1005186 (2016).
111. Sanborn, A. N. & Chater, N. Bayesian brains without probabilities. *Trends Cogn. Sci.* **20**, 883–893 (2016).
112. Dasgupta, I., Schulz, E., Goodman, N. & Gershman, S. Amortized hypothesis generation. Preprint at *bioRxiv* https://doi.org/10.1101/137190 (2017).
113. 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).
114. Gomez-Marin, A., Paton, J. J., Kampff, A. R., Costa, R. M. & Mainen, Z. F. Big behavioral data: psychology, ethology and the foundations of neuroscience. *Nat. Neurosci.* **17**, 1455–1462 (2014).
115. Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (MIT Press, Cambridge, MA, USA, 2010).
116. Love, B. C. The algorithmic level is the bridge between computation and brain. *Top. Cogn. Sci.* **7**, 230–242 (2015).
117. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. Preprint at https://arxiv.org/abs/1506.02142 (2016).
118. Rezende, D., Mohamed, S., Danihelka, I., Gregor, K. & Wierstra, D. One-shot generalization in deep generative models. *Proc. Int. Conf. Mach. Learn. Appl.* **48**, 1521–1529 (2016).
119. Kingma, D. & Welling, M. Auto-encoding variational Bayes. Preprint at https://arxiv.org/abs/1312.6114 (2013).
120. Naselaris, T. et al. Cognitive Computational Neuroscience: a new conference for an emerging discipline. *Trends Cogn. Sci.* **22**, 365–367 (2018).
121. Ahrens, M. B. et al. Brain-wide neuronal dynamics during motor adaptation in zebrafish. *Nature* **485**, 471–477 (2012).
122. Kietzmann, T., McClure, P. & Kriegeskorte, N. Deep neural networks in computational neuroscience. Preprint at bioRxiv https://doi.org/10.1101/133504 (2017).
123. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4**, 251–257 (1991).
124. Wyatte, D., Curran, T. & O'Reilly, R. The limits of feedforward vision: recurrent processing promotes robust object recognition when objects are degraded. *J. Cogn. Neurosci.* **24**, 2248–2261 (2012).
125. Spoerer, C. J., McClure, P. & Kriegeskorte, N. Recurrent convolutional neural networks: a better model of biological object recognition. *Front. Psychol.* **8**, 1551 (2017).
126. Hunt, L. T. & Hayden, B. Y. A distributed, hierarchical and recurrent framework for reward-based choice. *Nat. Rev. Neurosci.* **18**, 172–182 (2017).
127. Schäfer, A. M. & Zimmermann, H. G. Recurrent neural networks are universal approximators. *Int. J. Neural Syst.* **17**, 253–263 (2007).
128. O'Reilly, R. C., Hazy, T. E., Mollick, J., Mackie, P. & Herd, S. Goal-driven cognition in the brain: a computational framework. Preprint at http://arxiv.org/abs/1404.7591 (2014).
129. 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).

130. Schiess, M., Urbanczik, R. & Senn, W. Somato-dendritic synaptic plasticity and error-backpropagation in active dendrites. *PLoS Comput. Biol.* **12**, e1004638 (2016).
131. Marblestone, A. H., Wayne, G. & Kording, K. P. Towards an integration of deep learning and neuroscience. *Front. Comput. Neurosci.* **10**, 94 (2016).
132. Shadlen, M. N. & Shohamy, D. Decision making and sequential sampling from memory. *Neuron* **90**, 927–939 (2016).
133. Roelfsema, P. R. & van Ooyen, A. Attention-gated reinforcement learning of internal representations for classification. *Neural Comput.* **17**, 2176–2214 (2005).
134. Goodfellow, I. et al. Generative adversarial nets. Preprint at https://arxiv.org/abs/1406.2661 (2014).
135. Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A. & Hudspeth, A. J. *Principles of Neural Science* (McGraw-Hill Professional, New York, 2013).
136. Bastos, A. M. et al. Canonical microcircuits for predictive coding. *Neuron* **76**, 695–711 (2012).
137. Larkum, M. A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends Neurosci.* **36**, 141–151 (2013).
138. Fries, P. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends Cogn. Sci.* **9**, 474–480 (2005).
139. 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).
140. Yuille, A. & Kersten, D. Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* **10**, 301–308 (2006).
141. Helmholtz, H. *Handbuch der physiologischen Optik* (Dover, New York, 1860).
142. Gershman, S. J., Horvitz, E. J. & Tenenbaum, J. B. Computational rationality: a converging paradigm for intelligence in brains, minds, and machines. *Science* **349**, 273–278 (2015).
143. Simon, H. A. Bounded rationality. in *Utility and Probability* (eds. Eatwell, J., Milgate, M. & Newman, P.) 15–18, https://doi.org/10.1007/978-1-349-20568-4_5 (Palgrave Macmillan, London, 1990).
144. Griffiths, T. L., Lieder, F. & Goodman, N. D. Rational use of cognitive resources: levels of analysis between the computational and the algorithmic. *Top. Cogn. Sci.* **7**, 217–229 (2015).
145. Srikumar, V., Kundu, G. & Roth, D. On amortizing inference cost for structured prediction *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* 1114–1124 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2012).
146. Bengio, Y., Scellier, B., Bilaniuk, O., Sacramento, J. & Senn, W. Feedforward initialization for fast inference of deep generative networks is biologically plausible. Preprint at https://arxiv.org/abs/1606.01651 (2016).
147. Ghahramani, Z. Bayesian non-parametrics and the probabilistic approach to modelling. *Philos. Trans. A Math. Phys. Eng. Sci.* **371**, 20110553 (2012).
148. Deng, J. et al. ImageNet: a large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255, https://doi.org/10.1109/CVPR.2009.5206848 (IEEE, Piscataway, NJ, USA, 2009).
149. Beattie, C. et al. DeepMind Lab. Preprint at https://arxiv.org/abs/1612.03801 (2016).
150. Griffiths, T. L. Manifesto for a new (computational) cognitive revolution. *Cognition* **135**, 21–23 (2015).

Acknowledgements

This paper benefited from discussions in the context of the new conference Cognitive Computational Neuroscience, which had its inaugural meeting in New York City in September 2017[120]. We are grateful in particular to T. Naselaris, K. Kay, K. Kording, D. Shohamy, R. Poldrack, J. Diedrichsen, M. Bethge, R. Mok, T. Kietzmann, K. Storrs, M. Mur, T. Golan, M. Lengyel, M. Shadlen, D. Wolpert, A. Oliva, D. Yamins, J. Cohen, J. DiCarlo, T. Konkle, J. McDermott, N. Kanwisher, S. Gershman and J. Tenenbaum for inspiring discussions.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to N.K.

認知計算神経科学
Cognitive computational neuroscience

Nikolaus Kriegeskorte and Pamela K. Douglas

2018

出典: Nature Neuroscience, Vol. 1148(21), 2018, 1148–1160, www.nature.com/natureneuroscience

概要

認知が脳の中でどのように実現されているかを知るためには、認知タスクを実行できる計算モデルを構築し、そのモデルを脳や行動の実験で検証する必要がある。認知科学は、認知を機能的な要素に分解する計算モデルを開発してきた。計算神経科学では、相互作用するニューロンが認知の基本要素をどのように実現するかをモデル化している。今こそ、脳の計算というパズルのピースを組み立て、これらの別々の分野をよりよく統合する時である。現代の技術では、動物や人間の脳活動を、これまでにないほど豊富な方法で測定・操作することができる。しかし、このような実験は、脳計算モデルの検証に用いられて初めて、理論的な洞察を得ることができる。ここでは、認知科学、計算論的神経科学、人工知能の分野における最近の研究を紹介する。知覚・認知・制御タスクにおける脳の情報処理を模倣した計算モデルが開発され、脳や行動のデータを用いて検証され始めている。

脳の情報処理を理解するためには認知課題を実行できる計算モデルを構築する必要がある。課題を実行できる計算モデルを支持する論拠は 1973 年に Allen Newell が発表した「You can't play 20 questions with nature and win」という解説文によく現れている(1)。 Newell は認知心理学の現状を批判していた。認知心理学の分野では一連の二項対立の質問に自然に答えさせれば, やがて脳のアルゴリズムが明らかになると期待して, 認知に関する 1 つの仮説を一度に検証する習慣があった。 Newell は, 言葉で定義された認知に関する仮説を検証しても, 計算機的な理解にはつながらないと主張した。仮説を検証するためには, 課題を実行する包括的な計算モデルを構築することで補完する必要があると考えていた。認知機能を説明できるかどうか, 提案された構成要素のメカニズムが実際にどのような相互作用をしているのかは, コンピュータシミュレーションによる合成によってのみ明らかになる。もし, 情報処理メカニズムを完全に理解しているのであれば, それをエンジニアリングすることができるはずである。 1988 年に亡くなった物理学者のファインマンは「作れないものは分からない」という言葉を黒板に書き残している。

ここでは認知がどのようにして神経生物学的に妥当な動的要素から生じるのかを説明する, 課題実行可能な計算モデルが新しい認知計算神経科学の中心となることを主張する。まず認知科学と脳科学の歩みを簡単にたどり, 次に, 神経生物学的に妥当な人工知能 (AI) モデルを用いて認知科学 (人間がどのように学び, 考えるのかを説明する)(2) と計算神経科学 (脳がどのように適応し, 計算するのかを説明する) (3) の両方の野望を満たすことができるかもしれないことを示唆する, いくつかのエキサイティングな最近の進展をレビューする。

ニューウェルの批判精神に基づき, 認知心理学から認知科学への移行は, 課題を実行する計算モデルの導入によって定義された。認知科学者は認知を理解するには AI が必要であることを知っており，認知研究に工学を持ち込んだ。 1980 年代, 認知科学は記号的認知アーキテクチャ (4,5) とニューラルネットワーク(6) によって重要な進歩を遂げ, 人間の行動データを使って計算モデルの候補を判断するようになった。しかし, コンピュータのハードウェアと機械学習は, 認知プロセスの複雑さを完全にシミュレートするには十分ではなかった。さらに, これらの初期の開発では, 行動データだけに頼っていたため, 脳の解剖学的構造や活動から得られる制約を利用できなかった。

脳機能イメージングの登場により, 科学者たちは認知理論を人間の脳に関連づけるようになった。この試みは「認知神経科学」と呼ばれるようになった(7)。認知神経科学者は, 認知心理学の箱 (情報処理モジュール) と矢印 (モジュール間の相互作用) を脳にマッピングすることから始めた。これは脳の活動に関与するという点では前進した。だが, 計算の厳密さという点では後退した。認知科学の課題遂行型の計算モデルを, 脳活動データで検証する方法は考えられていなかった。その結果, 認知科学と認知神経科学は 1990 年代に決別することになった。

認知心理学が提唱する高レベルの機能モジュールの課題と理論は, 脳波計, ポジトロンエミッショントモグラフィー（PET), 初期の機能的磁気共鳴画像法 (fMRI) など, 空間分解能の低い機能画像技術を用いて, 人間の脳の粗いスケールの組織をマッピン

グするための合理的な出発点となった。認知心理学の「モジュール(8)」という概念にヒントを得て, 認知神経科学は「自然との 20 の質問」という独自のゲームを開発した。ある研究では, 脳の中に特定の認知モジュールが存在するかどうかが問われる。この分野では, 増え続ける認知機能を脳領域にマッピングし, 人間の脳の全体的な機能レイアウトの有用なラフな原稿を提供した。

どのようなスケールの脳地図であっても, 計算メカニズムを明らかにするものではない (図1)。しかし, マッピングは, 理論に制約を与える。結局のところ, 情報交換には, 通信する領域間の距離に応じて, 物理的な接続, エネルギー, 信号の待ち時間などのコストがかかる。部品の配置には, こうしたコストが反映されていると考えられる。高い帯域幅と短い応答時間で相互作用する必要のある領域は, 近くに配置されることが予想される(9)。より一般的には, 生物学的な神経ネットワークのトポロジーとジオメトリーは, そのダイナミクス, ひいては機能的なメカニズムを制約する。したがって, 機能的な局在化の結果は, 特に解剖学的な結合性と組み合わせることで, 最終的には脳の情報処理のモデル化に役立つと考えられる。

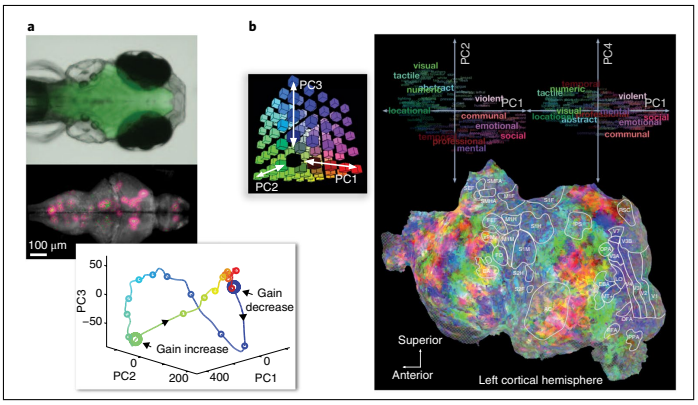


図1　現代のイメージング技術は, 脳の活動に関するかつてないほど詳細な情報を提供してくれるが, データに基づいた分析では限られた知見しか得られない。
a: 二光子カルシウムイメージングの結果 (121) は幼生のゼブラフィッシュが仮想環境と相互作用している間, 大規模な細胞集団の単一ニューロン活動を同時に測定したものである。 b: ヒトの fMRI の結果(70) は被験者が物語を聞いている間の意味選択的な反応の詳細なマップを示している。これらの研究は, 異なるスケールでの最新の脳活動測定技術の威力 (a,b) を示す一方で, このようなデータセットから脳の計算に関する洞察を引き出すことの難しさも示している。両研究とも, 複雑で時間的に連続した自然体験中の脳活動を測定し, 主成分分析 (a:下, b:上) を用いて, 活動パターンの全体像とその表現上の重要性を明らかにした。 (PC:主成分)。

方法論上の課題 (10,11) はあるものの, 認知神経科学で得られた知見の多くは, 確固たる基盤となる。例えばヒトの腹側視覚野に顔選択領域が存在するという知見 (12) は, 徹底的に再現され一般化されている。ヒト以外の霊長類でも, fMRI を用いて顔選択領域を調べたところ, 侵襲的な電極では広い視野で連続的な画像を得ることができないため, これまで調査できなかった同様の領域が見られた。霊長類の顔面パッチは fMRI で局在化され, 侵襲的な電極記録で検証された結果, 顔選択性ニューロンの密度が高いことが明らかになった(13)。また, 最前部のパッチでは, 鏡のように対称なチューニングや, 視野に依存しない個々の顔の表現など, 階層的な処理の高い段階で不変性が現れる(14)。顔認識の例は, 解剖学的基質のマッピングとニューロン応答の特性化が確実に進んでいることを示す一方で(15), 決定的な計算モデルがないことを示しているとも言える。ただし, 計算メカニズムの手がかりとなる文献もある。顔認識の脳-計算モデル (16) は顔選択ユニットの空間的クラスターや、fMRI (17,18) や侵襲的記録 (14,19) で観察される選択性や不変性を説明しなければならないだろう。

認知神経科学はヒトや霊長類の脳の全体的な機能配置を明らかにした(20)。しかし脳の情報処理を完全に計算で説明できるようにはなっていない。今後の課題は脳の構造や機能と一致し, 複雑な認知タスクを実行できる脳情報処理の計算モデルを構築することである。認知科学, 計算論的神経科学, 人工知能における以下のような最近の進展は, これが達成可能であることを示唆している。

1. 認知科学は, 複雑な認知プロセスを計算上の構成要素に分解することでトップダウンで進められてきた。認知科学は脳のデータを理解する必要性に悩まされることなく, 課題を実行する認知レベルの計算モデルを開発してきた。その成功例の一つが, ベイズ型認知モデルである。ベイズ型認知モデルは世界に関する事前知識と感覚的な証拠を最適に組み合わせることができる(21-23)。当初は基本的な感覚や運動のプロセスに適用されていた (23,24)。ベイズモデルは, 私たちの心が物理的・社会的世界をモデル化する方法など, 複雑な認知にも適用され始めている(2)。このような発展は, 統計学や機械学習との相互作用の中で起こり, 確率的な経験的推論に対する統一的な視点が生まれてきた。この文献は, 脳を理解するために不可欠な計算理論を提供し

ている。さらに, 実世界の知能に必要とされるような, 利用可能なデータに応じて複雑さを増すことができる生成モデルに対する近似的な推論のアルゴリズムも提供している (25,26)。

2. 計算論的神経科学は, 生物学的なニューロン間の動的な相互作用が, いかにして計算コンポーネントの機能を実現するかを示す, ボトムアップ的なアプローチをとってきた。この分野では, 過去 20 年間に初歩的な計算構成要素の数学的モデルを開発し, それらを生体ニューロンで実装してきた (27, 28)。この分野では, 感覚コーディング (29,30), 正規化 (31), 作業記憶 (32), 証拠の蓄積と決定メカニズム (33-35), 運動制御 (36) などのコンポーネントが含まれる。これらのコンポーネント機能のほとんどは, 計算上単純なものである。しかし, 認知のビルディングブロックを提供している。計算論的神経科学では, 高レベルの感覚や認知の脳内表現を説明できる複雑な計算モデルの検証も始まっている (37,38)。

3. 人工知能は, 構成要素の機能を組み合わせて知的な行動を作り出す方法を示した。初期の人工知能がその期待に応えられなかったのは, 知能を発揮するために必要な豊かな世界の知識を工学的に作ることも, 自動的に学習することもできなかったからである。最近の機械学習の進歩は, 計算能力の向上と学習対象となるデータセットの増加によって後押しされ, 知覚(39), 認知(40), 制御の課題において進歩をもたらした(41)。多くの進歩は, 認知レベルの記号的モデルによってもたらされた。最近の最も重要な進歩は, 入力の線形結合とそれに続く静的な非線形性を計算するユニットで構成されるディープニューラルネットワークモデルによるものである(42)。これらのモデルは活動電位などの基本的な機能を抽象化して, 生物学的なニューロンの動的機能のごく一部しか使用していない。しかし, これらの機能は脳にヒントを得たものであり, 生物学的なニューロンでも実装可能である。

これら 3 分野は, 認知課題を実行し, 脳の情報処理と行動を説明する生物学的に妥当な計算モデルに, 補完的な要素を提供する (図2)。ここでは認知科学 (認知課題を実行し, 行動を説明する計算モデル) と計算論的神経科学 (脳活動を説明する神経生物学的に妥当なメカニズムモデル) の成功基準を合わせて満たす認知計算論的神経科学に向けた文献の最初の一歩をレビューする。計算モデルが動物や人間の認知を説明するためには, 知能を発揮しなければならない。そのため AI 特に機械学習は認知的計算神経科学の理論的・技術的基盤となる重要な学問である。

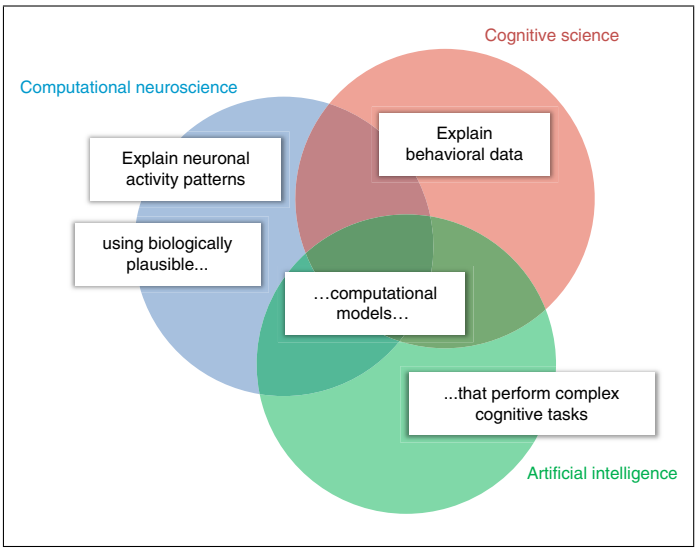


図2　脳の働きを理解するとはどういうことなのか？認知計算論的神経科学の目標は動物や人間の神経細胞の活動や行動の豊富な測定値を実世界の認知課題を実行する生物学的に妥当な計算モデルによって説明すること。歴史的に見て, 各分野 (丸) はこれらの課題 (白ラベル) のサブセットに取り組んできた。認知的計算論的神経科学はこれらすべての課題を同時に満たすことを目指している

そのためには, 理論 (課題を実行する計算モデルに具現化されている) と実験 (脳や行動のデータを提供している) の間にしっかりとした橋を架けることが重要な課題となる。このレビューの最初の部分では, 実験データから始まり, データから理論の方向に橋を架けようとするボトムアップ型の開発について説明する(43)。脳活動データがあれば, 接続性モデルは, 、脳の活性化の大規模なダイナミクスを明らかにすることを目的とし, 復号化・符号化モデルは, 脳の表現の内容と形式を明らかにすることを目的とする。この文献で採用されているモデルは, 計算理論に制約を与えるものではあるが, 一般に問題となっている認知タスクを実行するものではないので, 課題実行の基盤となっている計算メカニズムを説明するには至らない。

本稿の第 2 部では, 理論から実験への橋渡しをするという逆の方向に進む開発について説明する (37,38,44)。ここでは課題を実行する計算モデルを脳や行動のデータで検証し始めた新しい研究を紹介する。これらのモデルには, 抽象的な計算レベルで指

定され, 生物学的な脳への実装がまだ説明されていない認知モデルや, 神経生物学の多くの特徴を抽象化しているが生物学的なニューロンを使って実装することがもっともらしいニューラルネットワークモデルなどがある。これらの文献は, 脳の計算を理解するための統合的なアプローチの始まりを示唆している。すなわち認知課題を実行するためのモデルが必要であり, 生物学は許容できる構成要素の機能を提供し, 計算メカニズムは脳の活動や行動の詳細なパターンを説明するために最適化される。

1　実験から理論へ

■接続性と力学系のモデル Models of connectivity and dynamics . 脳の活動を測定して計算で理解する方法の一つとして、脳の接続性とダイナミクス(力学系) をモデル化する方法がある。接続性モデルは活性化された領域の局在を超えて領域間の相互作用を特徴づけるものである。ニューロンのダイナミクスは相互作用するニューロンの局所的なセットから脳全体の活動までさまざまなスケールで測定およびモデル化することができる(45)。脳のダイナミクスの一次近似は測定された応答時系列の相関行列によって得られ, これは場所間の対による「機能的連結性」を特徴づける。安静時ネットワークに関する文献では, このアプローチが検討されている(46)。空間独立成分分析のような時空間行列の線形分解でも, 同様に時間を超えた場所間の同時相関を捉えることができる(47)。

相関行列をしきい値にすることで, 地域の集合は無向グラフに変換され, グラフ理論的な方法で研究することができる。このような分析により, 「コミュニティ」(強く相互接続された領域の集合), 「ハブ」(他の多くの領域に接続された領域), 「リッチクラブ」(ハブのコミュニティ) が明らかになる(48)。接続性グラフは解剖学的または機能的な測定値から導き出すことができる。領域は解剖学的な経路を介して相互作用するため, 解剖学的な連結性グラフは機能的な連結性グラフによく似ている。しかし, 解剖学的結合性が機能的結合性を生み出す方法は, 局所的なダイナミクス, 遅延, 間接的な相互作用, ノイズを考慮に入れることで, よりよくモデル化される(49)。局所的な神経細胞の相互作用から, 大脳皮質や皮質下の領域にまたがる大規模な時空間パターンまで, 自発的なダイナミクスの生成モデルは, 脳活動データを用いて評価することができる。

有効連結性分析は, より仮説に基づいたアプローチをとり, ダイナミクスの生成モデルに基づいて小領域の集合間の相互作用を特徴づける(50)。活性化マッピングが認知心理学の箱を脳領域にマッピングするのに対し, 有効連結性分析は矢印を脳領域の対にマッピングする。この分野の研究のほとんどは, 脳領域の全体的な活性化のレベルで相互作用を特徴づけることに焦点を当てている。古典的な脳マッピング法と同様, これらの分析は領域平均活性化に基づいており, 領域間で交換される情報ではなく, 領域全体の活性化の相関的な変動を測定する。

有効連結性分析や大規模な脳ダイナミクスの分析は, 活性化や情報ベースのブレインマッピングで使用される線形モデルのような一般的な統計モデルを超えて, 生成モデルであるという点で, 測定値のレベルでデータを生成することができ, 脳のダイナミクスのモデルとなる。しかし, これらのモデルは, 表現された情報や, その情報が脳内でどのように処理されるかについては脳内でどのように処理されているかまでは把握できない。

■復号化モデル Decoding models　脳の計算メカニズムを理解するもう一つの方法は, 脳の各領域にどのような情報が存在するかを明らかにすることである。復号化はある課題に対するある領域の関与を示す活性化という概念を超えて, ある領域の集団活動に存在する情報を明らかにするのに役立つ。ある脳領域の活動から特定の内容が復号化可能であれば, それは情報の存在を意味する。脳領域が内容を「表現する」という言い方をすると, 情報がその信号を受け取る領域に内容を知らせる目的があるという機能的な解釈(51)が加わる。最終的にはこの解釈はその情報が他の領域や行動にどのような影響を与えるかについてのさらなる分析によって実証される必要がある(52-54)。

復号化モデルは神経細胞の記録に関する文献(27) にそのルーツがある。しかし, ニューロイメージング(55-59) では, 表彰の内容を研究するための一般的なツールとなっている。最も単純なケースでは, 復号化モデルによって, 2 つの刺激のうちどちらが測定された反応パターンを生じさせたかが明らかになる。表彰の内容は, 感覚刺激の同一性 (代替刺激のセットの中で認識されるべきもの), 刺激の特性 (格子の向きなど), 認知的な操作に必要な抽象的な変数, あるいは行動などである(60)。通常のように復号化モデルが線形である場合, 復号化可能な情報は, 下流のニューロンが 1 回のステップで読み出すことが可能な形式になっている。このような情報は, 活動パターンに「明示的」に含まれていると言われる(61)。

復号化モデルや他の種類の多変量パターン解析は, 脳の各領野の表彰内容を明らかにするのに役立ち(55,56,58,59), 脳=計算モデルが取り入れなければならない証拠となっている。しかし, 特定の情報をデコードする能力は, ニューロンコードを完全に説明するものではない。つまり, 表現形式 (線形デコード可能性を超えるもの) や, 他にどのような情報が存在するのかを特定するものではない。最も重要なことは, 復号化モデルは一般的に脳の計算モデルを構成するものではないということである。復号化モ

デルは, 脳の計算の過程ではなく, 生成物の側面を明らかにするものである。

■表彰モデル Representational models. 復号化モデルだけでなく, 私たちはある領域の表現を徹底的に特徴づけ, 任意の刺激に対する反応を説明したいと考えている。完全な特徴付けは, どのような変数がどの程度まで復号化できるかを定義することにもなる。表彰モデルは, 表彰空間を包括的に予測しようとするもので, 復号化モデルに比べて, 計算メカニズムに強い制約を与える(52,62)。

表彰モデル分析には, 符号化モデル(63-65), パターン成分モデル(66), 表彰類似性分析(57,67,68) の 3 種類が文献に紹介されている。これら 3 つの方法はいずれも, 実験条件の多変量記述 - 例えば, 刺激のセットの意味的な記述や, 刺激を処理するニューラルネットワークモデルの層全体の活動パターン-に基づいて, 表彰空間に関する仮説を検証するものである(52)。

符号化モデルでは, 刺激に対する各ボクセルの活動プロファイルは, モデルの特徴の線形結合として予測される。パターン成分モデルでは, 表彰空間を特徴づける活動プロファイルの分布を多変量正規分布としてモデル化する。表象類似性分析では, 刺激によって誘発される活動パターンの表現上の非類似性によって表現空間を特徴づける。

表彰モデルは, 人間の観察者が提供するラベルのような, 刺激の記述に基づいて定義されることが多い(63,69,70)。このシナリオでは, ある領域における脳の反応を説明する表象モデルは, 脳の計算論的説明ではなく, 少なくとも表象の記述的説明を提供することになる。このような説明は, モデルが新しい刺激に一般化するとき, 計算理論への有効な足がかりとなる。重要なことは, 表象モデルは脳=計算モデル間の判断を可能にすることである。

本節では, 脳活動データから計算上の知見を得るのに役立つ 3 種類のモデルについて考察した。接続性モデルは, 領域間の動的な相互作用の側面を捉える。復号化モデルは, 脳の領域を調べてその表現内容を明らかにすることができる。表象モデルは, ある領域の表現空間を完全に特徴づける明確な仮説を検証することができる。これら 3 種類のモデルはいずれも, 理論的に動機づけられた疑問を解決するために用いることができる。しかし, 課題を実行できる計算モデルがない場合, 一連の質問をしても, 説明しようとしている認知機能の根底にある計算メカニズムを明らかにすることはできないという Newell の議論に従うことになる。これらの方法は, 認知機能の基礎となる情報処理がどのように機能しているかを具体的に示すメカニズムモデルを検証しないため, 理論への橋渡しにはならない。

2 理論から実験へ From theory to experiment

実験と理論の間により良い橋を架けるためには, まず理論を完全に特定する必要がある。そのためには, 理論を数学的に定義し, それを計算モデルで実装する必要がある (Box 1)。計算モデルは認知的な忠実さと生物学的な忠実さを両立させるためにさまざまなレベルの記述を行うことができる (図3)。ニューロンの構成要素とダイナミクスのみを捉えるように設計されたモデル(71)は, 認知機能 (72) を説明できない傾向がある (図3横軸)。逆に認知機能だけを捉えて設計されたモデルは脳との関連付けが難しい (図3縦軸)。心と脳を結びつけるためには行動と神経細胞のダイナミクスの両方の側面を捉えようとするモデルが必要である。最近の研究では脳からの制約が認知機能の説明に役立つことが示唆されており(42,73,74), その逆もまた同様で(37,38), トレードオフを統合に変えることができる。

Box 1.「モデル」の多義性

「モデル」という言葉は脳科学や行動科学において多くの意味を持っている。データ解析モデルは測定された変数間の関係を確立するための一般的な統計モデルである。例えば線形相関ブレインマッピングのための一変量重回帰, 線形デコーディング分析などがある。「効率接続性モデル」や「因果関係モデル」も同様, データ分析モデルである。これらのモデルは脳領域間の因果関係や相互作用を推測するのに役立つ。データ解析モデルは, 変数間の関係 (例えば, 相関, 情報, 因果関係) に関する仮説を検証する目的で使用することができる。脳の情報処理のモデルではない。一方, 「箱と矢印モデル」は認知要素の機能を表すラベル付きの箱と情報の流れを表す矢印で構成された情報処理モデルである。認知心理学の分野では, 脳の計算理論を解明するために, 定義されていないにもかかわらず, このようなモデルが有用であった。「言葉モデル」も同様に, 脳の情報処理に関する理論を, 言葉で曖昧に定義したスケッチである。これらは情報処理のモデルではあるが, 脳で行われていると考えられている情報処理を実行するものではない。「オラクル神託モデル」とは脳の反応をモデル化したもの (データ解析モデルに実体化されていることが多い) で脳をモデル化している動物が入手できない情報に依存しているものである。例えば腹側頭の視覚反応を抽象的な形状の記述の関数としてあるいはカテゴリーラベルや連続的な意味的特徴の関数としてモデル化した場合, そのモデルが画像から形状, カテゴリー, 意味的特徴を計算することができなければオラクル神託モデルとなる。オラクルモデルは、ある領域に存在する情報とその表現形式の有用な特性を提供することができるが、その表現が脳によってどのように計算されるかについては、い

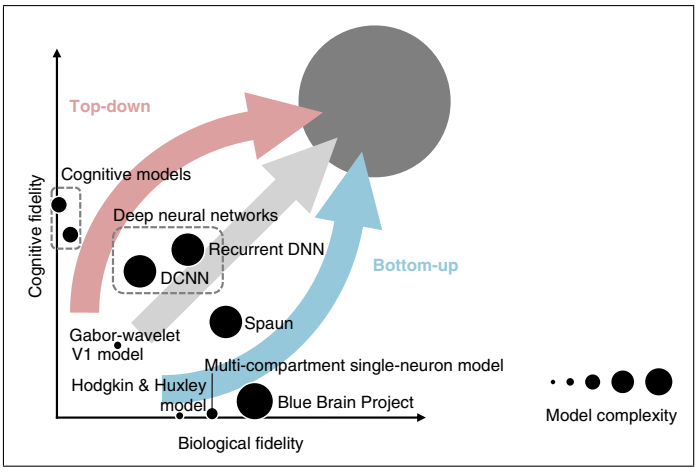


図3　脳内で起こるプロセスのモデルは、さまざまなレベルの記述で定義することができ、そのパラメトリックな複雑さ（ドットサイズ）や、生物学的（横軸）および認知的（縦軸）な忠実度もさまざまです。理論家は、さまざまな主要目標を掲げてモデリングに取り組んでいます。ボトムアップ型のモデリング（青矢印）は、まず、活動電位や単一ニューロンの複数のコンパートメント間の相互作用など、生物学的な神経ネットワークの特性を把握することを目的としています。このアプローチでは、認知機能を無視して、皮質の柱や領域などの脳の小さな部分の創発的なダイナミクスを理解し、振動などの生物学的なネットワーク現象を再現することに集中します。振動などの生物学的ネットワーク現象を再現する。トップダウンアプローチ（赤矢印）は、まず認知機能をアルゴリズムレベルで捉えることを目的としています。これは、生物学的な実装を無視して、タスクパフォーマンスの基礎となる情報処理をアルゴリズムの構成要素に分解することに焦点を当てたアプローチである。この2つのアプローチは、「脳がどのようにして心を生み出しているのか」という共通の目標に向かって、連続した道の両極を形成しています。全体として、認知的忠実度と生物学的忠実度の間にはトレードオフ（負の相関関係）がある。しかし、認知的な制約が生物学的な機能を明らかにし、生物学が認知的な偉業を説明するモデルを刺激することで、トレードオフは相乗効果（正の相関）に変わります。知能には豊かな世界観が必要なので、人間の脳の情報処理のモデルは、パラメトリックな複雑さが高くなります（右上の大きな点）。生物学的な詳細を排除したモデルでタスクパフォーマンスを説明できたとしても、神経生物学的な実装を説明するためには、生物学的に詳細なモデルが必要になります。この図は、モデル間の関係を理解し、それぞれの補完的な貢献を評価するのに役立つ概念図です。しかし、この図は、認知的忠実度、生物学的忠実度、モデルの複雑さを定量的に測定したものではありません。この3つの変数をそれぞれ測定する明確な方法はまだ開発されていません。図は参考文献 (122)

かなる理論も規定しない。一方「脳=計算モデル (BCM)」はある課題を実行する際の脳の情報処理を, ある程度の抽象度で模倣したモデルである。例えば, 視覚神経科学の分野では, 画像のビットマップを入力として, 脳の活動や行動を予測する視覚処理の BCM が画像計算可能モデルと呼ばれている。ディープニューラルネットは画像計算可能な視覚処理のモデルを提供する。しかし, スーパービジョンによって学習されたディープニューラルネットは, 学習のためにカテゴリーラベル付きの画像に依存している。生物の発達や学習の過程では, ラベル付きの例は (同等の量で) 入手できないため, これらのモデルは視覚処理の BCM ではあるが, 発達や学習の BCM ではない。「強化学習モデル」はより現実的な質の環境フィードバックを用いるので, 学習過程の BCM となりうる。「感覚符号化モデル」は感覚入力をある段階の内部表現に変換する計算の BCM である。「内部変換モデル」は 2 つの処理段階の間で表現が変換されることの BCM である。「行動復号モデル」とはある内部表現から行動出力への変換の BCM である。 BCM というラベルは単にそのモデルがある程度の抽象度で脳の計算を捉えることを意図していることを示していることに注意。 BCM は生物学的な詳細から任意の程度まで抽象化することができるが, し脳の活動や行動の何らかの側面を予測しなければならない。感覚入力から行動出力を予測する「心理物理学モデル」や認知タスクを実行する「認知モデル」は, 高いレベルの記述で定式化された BCM である。 BCM というラベルは, そのモデルがもっともらしいとか, 経験的なデータと一致するとかいうことを意味するものではない。 BCM の候補を経験的に否定することで進歩する。 BCM は, 脳の計算を支える生物学的プロセスを表現するミクロスケールの「生物物理モデル」や, マクロスケールの「ブレインダイナミカルモデル」や「因果インタラクションモデル」と同様に，脳で発生するプロセスのモデルである．しかし他のプロセスモデルとは異なり BCM は脳のダイナミクスの機能と考えられている情報処理を行う。最後にモデルベース強化学習やモデルベース認知のように脳が採用する世界のモデルを指す場合には「モデル」という言葉が使われる。

2.1 Neural network models

ニューラルネットワークモデル（囲み記事 2）の歴史は長く、様々な分野で織り交ぜられています。計算論的神経科学では、様々なレベルの生物学的詳細に基づくニューラルネットワークモデルが、生物学的神経ネットワークのダイナミクスや初歩的な計算機能を理解するために不可欠なものとなっています (27,28)。認知科学においては、1980 年代に並列分散処理と呼ばれる認知機能を理解するための新しいパラダイムを定義し (6,75), この分野を神経科学に近づけました。 AI においては、視覚や音声認識などの知覚タスクから、言語翻訳などの記号処理タスク、さらには音声合成やロボット制御などの運動タスクに至るまで、多くのアプリケーションに大きな進歩をもたらしました (42,74)。ニューラルネットワークモデルは、3 つの分野の成功基準を合わせて満たす、タスクパフォーマンスの高いモデルを構築するための共通言語です（図2)。

ニューラルネットワークモデルは、脳と同様に、フィードフォワード計算とリカレント計算を行うことができます (37,76)。最近の進歩を支えているモデルは，線形-非線形の信号変換を複数の段階で構成するという意味で，深いものです．モデルは通常、数百万のパラメータ（接続重み）を持ち、タスクのパフォーマンスを最適化するように設定されます。成功したパラダイムの一つは教師付き学習で、入力（画像など）と関連する出力（カテゴリーラベルなど）のトレーニングセットから、入力から出力への望ましいマッピングを学習します。しかし、ニューラルネットワークモデルは、教師なしで学習することもでき、経験データに固有の複雑な統計的構造を学習することができる。

パラメータの数が多いと、解釈可能なパラメータの数が少ない単純なモデルに慣れている研究者は不安になります。しかし、単純なモデルでは、複雑な知能の働きを説明することはできません。 AIの歴史は、知能には十分な世界の知識と、それを蓄えるための十分なパラメータの複雑さが必要であることを示している。そのため、私たちは複雑なモデル（図3）と、それがもたらす課題に取り組まなければなりません。一つの課題は、パラメータ数が多いため、モデルを理解するのが難しいことです。モデルは完全に透明であるため、内部表現を理解するために何百万もの入力パターンを安価に調べることができます。オーバーフィッティングの懸念に対処するため、モデルはその一般化性能の観点から評価されます。例えば、視覚モデルは、学習していない画像に対する神経活動や行動反応を予測する能力の観点から評価されます。

囲み記事2: ニューラルネットワークモデル **ニューラルネットワークモデル** とは、生物学的なニューラルネットワークにヒントを得て開発されたモデルで、各ユニットが多数の入力を組み合わせ、情報がネットワークを通じて並行して処理されるという特徴を持っています。生物学的に詳細なモデルでは、活動電位や各ニューロンの複数のコンパートメントにおけるダイナミクスを捉えることができますが、ニューラルネットワークモデルでは、生物学的な詳細は除外されています。しかし、視覚的な物体認識などの認知機能を説明することができるため、認知と脳を結びつけるための魅力的なフレームワークとなっている。
典型的なユニットは、入力の線形結合を計算し、その結果を静的な非線形性に通します。その出力は、ニューロンの発火率に似ていると解釈されることもある。浅いネットワーク（入力と出力の間に隠れユニットの層が1つあるもの）でも、任意の関数を近似することができます[123]。しかし，深層ネットワーク（複数の隠れ層を持つネットワーク）は，実世界のタスクで必要とされる複雑な機能の多くをより効率的に捉えることができます．コンピュータビジョンをはじめとする多くのアプリケーションでは，**¿フィードフォワード** アーキテクチャが使用されている．しかし，ユニットの出力を再処理して複雑なダイナミクスを生成する ¡b¿リカレントニューラルネットワーク¡/b¿ は，さらなる工学的進歩をもたらし[76]，脳内の再帰的なシグナル伝達をよりよく表現している[35,124-126]．フィードフォワードネットワークが **普遍関数近似器** であるのに対し、リカレントネットワークは **動的システムの普遍的な近似器** である(127)。リカレント処理は、ネットワークがその限られた計算資源を時間の経過とともに再利用し、より複雑な計算のシーケンスを実行することを可能にする。リカレントネットワークは、最近の刺激の履歴を動的に圧縮して表現し、現在の処理に必要な時間的コンテキスト情報を提供することができる。その結果、リカレントネットワークは動的なパターンを認識し、予測し、生成することができる。
フィードフォワードネットワークもリカレントネットワークも、そのアーキテクチャと接続の重みの設定によって定義される。重みを設定する方法の一つとして、出力をある目的の出力に近づけるために小さな調整を繰り返す方法があります (**教師付き学習**)。各重みは、それを少し変更することで得られる誤差の減少に比例して調整されます。この方法は，誤差が最も急峻に減少するような重みの空間のステップを生成することから，**勾配降下法** と呼ばれる。勾配降下法は、各重みに対する誤差関数の微分を計算する効率的なアルゴリズムである **バックプロパゲーション** を用いて実装することができる。
脳がバックプロパゲーションのようなアルゴリズムを使って学習しているかどうかは議論の余地がある。バックプロパゲーションやそれに近い形の教師付き学習の生物学的に妥当な実装方法がいくつか提案されている (128-130)。 ¡b¿教師信号¡/b¿ は、複

数の感覚モダリティが提供する文脈に基づいて内部的に生成されるかもしれない[131]。感覚や記憶からより多くの証拠が得られるようになり、時間の経過とともに表現が動的に洗練されていくことに基づいて (132), 環境との相互作用で生じる内部および外部の強化信号に基づいて (133)。 **強化学習** (41) や, ニューラルネットワークのパラメータの **教師なし学習** (119,134) は, 現在急速に進歩している分野です。ニューラルネットワークのモデルは、生物学からヒントを得ることで、AI に飛躍的な進歩をもたらすことを実証しています。人間の認知能力に匹敵するモデルを求めて、生物学に深く入り込んでいくことになりそうです[135]。現在、工学的に最も成功している抽象的なニューラルネットワークモデルは、生物学的なハードウェアでも実装可能です。しかし、これらのモデルは、脳の動的な構成要素のごく一部しか使用していません。神経科学では、活動電位(108), 正規の微小回路(136), 樹状突起のダイナミクス (128,130,137), 振動 (138) などのネットワーク現象 (27) など、計算機能を持つ可能性のある動的コンポーネントの豊富なレパートリーが記述されています。また、生物学は、グローバルなアーキテクチャに制約を与え、例えば、学習のための補完的なサブシステムを示唆しています139。これらの生物学的要素を、意味のあるタスクを実行するように設計されたニューラルネットワークの文脈でモデル化することで、脳の計算にどのように貢献しているかが明らかになり、AI のさらなる発展につながるかもしれません。

最近のいくつかの研究では，ニューラルネットワークモデルを脳の情報処理のモデルとして検証し始めています[37,38]．これらの研究では，画像中の物体を認識するように訓練された深層畳み込みニューラルネットワークのモデルを用いて，霊長類の腹側視覚野における新規画像の脳内表現を予測した．モデルを用いて予測しました．その結果、深層畳み込みニューラルネットワークの内部表現は、ヒトとサルの下側頭葉皮質における視覚イメージの表現について、現在最も優れたモデルを提供することが明らかになった[77-79]。多数のモデルを比較した場合，物体の分類というタスクを実行するために最適化されたモデルが，皮質の表現をよりよく説明していた[77,78].

物体を認識するように訓練された深層ニューラルネットワークの初期の層は、初期の視覚野の表現に似たものを含んでいる[78,80]。腹側視覚ストリームに沿って移動すると、ニューラルネットワークのより高い層が、表現を説明するためのより良い基盤を提供するようになる[80-82]。深い畳み込み神経回路網の高層は、物体の位置、大きさ、姿勢、そして物体のカテゴリーを復号することができるという点で、下側頭皮質の表現にも似ています[83]。この分野では、脳の活動データを予測することでこれらのモデルを検証することに加えて、知覚された形状84や物体の類似性を反映した行動反応を予測することでモデルを検証し始めている[85]。

2.2　認知モデル

認知レベルのモデルでは、神経生物学的に妥当な構成要素を用いてその実装に取り組むことなく、情報処理のイメージを描くことができます。これにより、ニューラルネットワークモデルではまだ不十分な高次認知領域の研究を進めることができます。さらに、あるプロセスがニューラルネットワークモデルでも実現可能な場合でも、認知モデルが有用な抽象化を提供することもあります。

現在、脳科学的な説明は、感覚や運動のプロセスが動物と環境を結びつける脳の周辺部に近い機能的な構成要素を支配しています。しかし、高次の認知機能の多くは、脳科学的な説明やニューラルネットワークモデルでは理解できないものでした。認知モデルのユニークな貢献を説明するために、生産システム、強化学習モデル、ベイズ認知モデルという3つのクラスの認知モデルについて簡単に説明する。

プロダクションシステムは、推論や問題解決を説明できる認知モデルのクラスの初期の例です。これらのモデルは、ルールとロジックを使用しており、感覚データや運動信号ではなくシンボルで動作するという点でシンボリックである。これらのモデルは、認知を物理的環境に基づかせる知覚や運動制御ではなく、認知を捉えます。プロダクション」とは、「if-thenルール」に基づいて行われる認知的な行動のことである。このようなルールのセットは、一連の生産物（「then」）が実行される条件（「if」）を規定する。条件とは、現在の目標や記憶にある知識のことである。アクションは、目標や知識の内部状態を変更することができる。例えば、プロダクションはサブゴールを作成したり、推論を保存したりすることができる。複数のルールで条件が満たされた場合、競合解決メカニズムが1つのプロダクションを選択します。この形式論を用いて指定されたモデルは、一連のプロダクションを生成します。これは、ある認知的な目標に向かって作業をしている私たちの意識的な思考の流れに、ある程度似ているかもしれません。また、プロダクション・システムの形式論は、普遍的な計算アーキテクチャーを提供するものでもある[86]。ACT-R[5]のようなプロダクション・システムは、もともと行動データの指導のもとに開発されたものである。最近では、このようなモデルは、地域平均のfMRI活性化タイムコースを予測する能力という点でもテストされ始めている (87)。

強化学習モデルは、エージェントが環境との相互作用を通じて、長期的な累積報酬を最大化するように学習する方法を示しています[88,89]。生産システムと同様に，強化学習モデルでは，エージェントが状態と行動の離散的な記号表現を使用できる知覚モジュールと運動モジュールを備えていることを前提としています。エージェントは行動を選択し、その結果としての環境の状態を観察し、途中で報酬を受け取り、行動を改善するように学習します。エージェントは、各状態とその期待される累積報酬を関連付ける「価値関数」を学習することができる。エージェントは、各行動がどの状態につながるかを予測し、それらの状態の値を知っていれば、最も有望な行動を選択することができます。エージェントは、各状態と有望なアクションを直接関連付ける「ポリシー」を学習することもできます。行動の選択は、搾取（短期的な報酬をもたらす）と探索（学習に利益をもたらし、長期的な報酬をもたらす）のバランスをとる必要がある。

強化学習の分野では、累積報酬を最大化するためにどのように行動し、学習するかを定義するアルゴリズムを探求します。強化学習理論は、心理学や神経科学にルーツを持ち、現在では機械学習やAIの重要な分野となっています。強化学習理論は、古典的な手法である動的計画法、モンテカルロ法、網羅的探索法を限界事例として含む、非常に一般的な制御の視点を提供し、環境が確率的で部分的にしか観測されず、その因果関係のメカニズムが不明であるような困難なシナリオを扱うことができます。

エージェントは、環境を徹底的に探索し、どのような状態でも最も有望な行動を試行錯誤で学ぶことができるかもしれません（モデルフリー制御)。そのためには、十分な学習時間と記憶力、そしてエージェントが早死にしないような環境が必要です。しかし、生物は学習時間や記憶力が限られており、死に至るような相互作用を避けなければなりません。このような状況では、エージェントは環境のモデルを構築した方が良いかもしれません。モデルは、経験を圧縮して一般化し、新しい状況での知的行動を可能にする（モデルベース制御)。モデルを用いない方法は，計算効率が高い（状態から値へのマッピング，あるいは行動への直接のマッピング）が，統計的には効率が悪い（学習に時間がかかる)．モデルベースの方法は，統計的には効率が良いが，（起こりうる未来をシミュレートするために）膨大な計算量を必要とする場合がある (90)。

信頼できるモデルを構築するのに十分な経験が得られるまでは、エージェントは単にエピソードを保存し、過去に成功した行動経路に戻るのが最善かもしれない（エピソード制御）[91,92]。エピソードを保存することで、モデル構築に重要な逐次的な依存関係の情報が保存される。さらに、エピソード制御は、成功した行動経路を支える因果関係を理解する前に、そのような依存性を利用することができる。

脳は、これらの3つの制御モード（モデルフリー、モデルベース、エピソード）をそれぞれ行うことができ (89), まだ発見されていないアルゴリズムを用いて, それぞれの利点を組み合わせているように見える。 AI と計算論的神経科学は、このアルゴリズムを発見するという目標を共有しています (41, 90, 93-95) が、この目標に対して異なる角度からアプローチしています。これは、認知的な課題が形式モデルの開発の動機となり、AIと神経科学の進歩を促す例です。

認知モデルの第三のクラスは、ベイズモデル（囲み記事 3）である[21,96-98]。ベイズ推論は、認知についての本質的な規範的視点を提供する。動物が最適な行動をとるために、脳が実際に何を計算すべきかを教えてくれる。例えば、知覚的な推論では、現在の感覚データを事前の信念の文脈の中で考慮する必要がある。ベイジアン推論とは、確率のルールに従って、データと事前の信念を組み合わせることです。

囲み記事3: ベイズ認知モデルベイズ認知モデルは、「脳はある課題に対して統計的に最適な解を近似的に求める」という仮定に基づいています。推論を行い、何をすべきかを決定するための統計的に最適な方法は、確率の法則を用いて、現在の感覚的な証拠をすべての利用可能な事前知識に照らして解釈することである。統計的に最適な方法は、現在の感覚的な証拠を、確率の法則を用いて、利用可能なすべての予備知識に照らして解釈することです。視覚の場合を考えてみましょう。網膜の信号は、私たちが認識したいと思っている世界のオブジェクトを反映しています。物体を推論するためには、どのような物体の構成が考えられるか、それぞれがどの程度画像を説明できるかを考える必要があります。我々の事前の信念は、オブジェクトの各構成の確率と、与えられた構成が異なる網膜画像を生成する確率を把握する生成モデルによって表されます。
より形式的には、ベイズ視覚モデルは、感覚データ d（画像）と世界の原因 c（推論すべき表面、物体、光源の構成）の結合分布 $p(d, c)$ の生成モデルを使用することになる[140]。同時分布 $p(d, c)$ は、原因のすべての可能な構成に対する事前分布 $p(c)$ と尤度の積に等しい。原因の特定の構成が与えられたときの、特定の画像の確率である $p(d|c)$ と等しい。 $p(d|c)$ の所定のモデルがあれば、特定の原因 c から特定の画像 d が生まれる確率である尤度を評価することができます。また、$p(d|c)$ の暗黙のモデルとして、原因 c からデータ d（画像）への確率的なマッピングがある場合もあります。このようなモデルは、自然な画像を生成します。 $p(d|c)$ のモデルは、規定されたものであれ、暗黙のものであれ、世界の原因がどのように画像を作り出すか、少なくとも画像とどのように関連するかを捉えています。視覚認識は、ある画像が与えられたときの原因の確率分布である事後分布$p(c|d)$eを計算することになります。事後 $p(c|d)$ eは、感覚データdを説明するために世界に存在しなければならない原因 c

を明らかにする[141]。 $p(c|d)$ を計算するモデルは、画像を判別することから、判別モデルと呼ばれています（ここでは、効果（画像）から原因へのマッピング）。反転には、数学的には、潜在的な原因に対する事前情報 $p(c)$ が必要です。事前の $p(c)$ は解釈を制約し、どのような画像も説明できる原因の複数の構成から生じる曖昧さを軽減するのに役立ちます。

原因 c の推論を、利用可能なすべての知識と不確実性を取り込んだ $p(d, c)$ の生成モデルに基づいて行うことは、統計的には最適（限られたデータで最良の推論が得られる）であるが、計算上は困難（動物が使用できる以上のニューロンや時間が必要になる可能性がある）である。理想的には、推論$p(c|d)$に暗黙的に含まれる生成モデル$p(d, c)$は、画像形成に関する知識だけでなく、世界にあるものやdに関する知識も含んでいなければなりません。推論$p(d, c)$に暗黙的に含まれる生成モデルは、画像形成に関する知識だけでなく、世界の物事やそれらの相互作用、そしてこれらのプロセスに関する我々の不確実性も含んでいなければならない。一つの課題は、感覚データから生成モデルを学習することです。その際には、学習した知識と残りの不確実性を表現する必要があります。もし、生成モデルの仕様が間違っていれば、推論は最適なものにはなりません。現実のタスクでは、ある程度のモデルの誤指定は避けられません。例えば、生成モデルには、画像生成プロセスの過度に単純化されたバージョンが含まれている可能性があります。もう一つの課題は、事後評価 $p(c|d)$ の計算である。現実的に複雑な生成モデルでは、マルコフ連鎖モンテカルロ法、信念伝播法、変分法など、計算量の多い反復アルゴリズムを用いた推論が必要になることがあります。統計的な効率と計算的な効率の間で脳がとる妥協[142-144]は、高速なフィードフォワード認識モデルを学習することで、頻出する成分の推論を高速化し、反復アルゴリズムでは流動的に導出できない結論を結晶化することであろう。これは amortized inference[145,146] として知られている。

ベイズ型認知モデルは、最近、機械学習や統計学との相互作用の中で発展してきた。初期の研究では、固定された構造を持つ生成モデルを使用しており、限られたパラメータのセットに関してのみ柔軟性があった。現代の生成モデルは、データとともに複雑さを増し、固有の構造を発見することができる[98]。これらのモデルは、あらかじめ定義された有限のパラメータセットによって制限されないため、ノンパラメトリックと呼ばれている[147]。それらのパラメータは、事前に定義された境界なしに数を増やすことができる。

ベイズモデルは、基本的な感覚や運動のプロセスの理解に貢献してきた[22-24]。また、古典的な認知バイアス[99]を、実験課題では間違っていても、現実世界では正しく役立つ事前の仮定の産物として説明することで、判断や意思決定といった高次の認知プロセスに対する洞察を提供してきた。

ベイズ型ノンパラメトリックモデルにより、認知科学はより複雑な認知能力を説明し始めている。例えば、1つの例から新しい物体カテゴリーを誘導する人間の能力を考えてみましょう。このような帰納的な推論には、現在のフィードフォワード・ニューラルネットワークモデルでは捉えられない種類の事前知識が必要である[100]。カテゴリーを誘導するためには，対象物，その部品間の相互作用，それらがどのようにしてその機能を生み出すかについての理解に頼ることになる．ベイズ認知の視点では、人間の心は幼児期から世界のメンタルモデルを構築します[2]。これらのモデルは、確率的な意味での生成モデルであるだけでなく、因果的で構成的である場合もあり、新規のシナリオや仮説的なシナリオに一般化するために再構成可能な要素を用いて、世界のプロセスの精神的なシミュレーションをサポートします[2,98,101]。このモデリング・アプローチは、物理的世界[101-103]、さらには社会的世界[104]についての推論にも適用されている。

生成モデルは、一般的な知能に不可欠な要素である。生成モデルを学習しようとするエージェントは、その経験の間のすべての関係を理解しようとする。学習のために外部からの監視や強化を必要とせず、環境や自分自身についての洞察を得るためにすべての経験を掘り起こすことができる。特に、世界のプロセスの因果モデル（物体がどのように画像を引き起こすのか、現在がどのように未来を引き起こすのか）は、エージェントに深い理解を与え、推論や行動のより良い基盤となります。

ニューロン集団における確率分布の表現については、理論的にも実験的にも検討されてきた[105,106]。しかし、ベイズ推論や学習、特にノンパラメトリックモデルでの構造学習を、脳内での実装に関連付けることは、依然として困難である[107]。脳の計算理論として、サンプリングなどの近似的な推論アルゴリズムは、皮質のフィードバック信号や活動の相関を説明できるかもしれない[97,108-110]。さらに、計算効率を上げるために脳が削減する手抜き、つまり近似性が、人間の統計的最適性からの逸脱を説明する可能性もある。特に、認知実験では、人間の行動にサンプリング[111]や償却型推論[112]の特徴があることが明らかになっている。

ここで取り上げた3つのクラスを含む認知モデルは、認知を意味のある機能的な構成要素に分解する。認知科学者は、脳内の実装から独立したモデルを宣言することで、現在のニューラルネットワークでは実現できない高レベルの認知プロセス[21,97,98]を扱うことができる。認知モデルは、部分の役割を理解しようとするときに全体を見ることができるため、認知的計算論的神経科学には欠かせないものです。

囲み記事4: 認知科学, 計算論的神経科学, AI はなぜお互いに必要なのか？

認知科学 は単に認知モデルの脳への実装を説明するだけでなく、そのアルゴリズムを発見するために、計算論的神経科学を必要としています。例えば、感覚処理や物体認識の主要なモデルは、脳にインスパイアされたニューラルネットワークですが、その計算は認知レベルでは簡単には捉えられません。また、最近のベイズノンパラメトリックモデルの成功は、一般的にはまだ実世界の認知には適用できません。人間の認知機能の計算効率を説明し、認知機能の詳細なダイナミクスや行動を予測するには、脳活動のダイナミクスを研究することが有効である。行動を説明することは重要であるが、行動データだけでは複雑なモデルの制約条件としては不十分である。しかし、脳のデータを適切に活用すれば、認知アルゴリズムに豊富な制約を与えることができます。認知科学は、常に人工知能と密接な関係を保ちながら発展してきました。両分野は、タスクを実行するモデルを構築するという目的を共有しており、共通の数学的理論とプログラミング環境を利用しています。

計算論的神経科学 は、より高度な認知に挑戦するために認知科学を必要としている。実験レベルでは、認知科学の課題により、計算論的神経科学が認知を実験室に持ち込むことができます。理論レベルでは、認知科学は、計算神経科学が研究している神経生物学的な動的構成要素が、認知や行動にどのように寄与しているかを説明することに挑戦します。計算論的神経科学は、生物学的に妥当な動的要素を持つ認知機能をモデル化するための理論的・技術的基盤を提供するために、AI、特に機械学習を必要としています。認知機能を生物学的に妥当な動的要素でモデル化するための理論的・技術的基盤を提供するために、AI、特に機械学習が必要です。¡br/¿ **人工知能** を実現するためには、認知科学が必要である。認知科学の課題は AI システムのベンチマークとなり、初歩的な認知能力から人工的な一般知能へと積み上げていくことができます。人間の発達と学習に関する文献は、学習者が何を達成することが可能か、また、どのような種類の世界との相互作用が知能の獲得をサポートするかについての重要な指針となります。人工知能には、アルゴリズムのヒントとなる計算論的神経科学が必要です。ニューラルネットワークモデルは、脳からヒントを得た技術の一例であり、AI のいくつかの分野で他の追随を許さないものです。神経生物学的な動的構成要素（例えば、スパイクニューロン、樹状突起のダイナミクス、皮質の典型的な微小回路、振動、神経調節プロセスなど）や、人間の脳のグローバルな機能レイアウト（例えば、感覚モダリティ、記憶、計画、運動制御などの異なる機能に特化したサブシステムなど）からさらなるインスピレーションを得ることができれば、AIのさらなるブレークスルーにつながるかもしれません。機械学習は、統計学と計算機科学という別々の伝統に基づいており、それぞれ統計的な効率と計算的な効率を最適化してきました。計算効率と統計効率の統合は、ビッグデータ時代の必須課題です。脳は計算効率と統計効率の両方を兼ね備えていると考えられ、そのアルゴリズムを理解することで機械学習を促進できるかもしれません。

囲み記事5: 共有可能な課題, データ, モデル, テスト:学際的なコラボレーションの新しい文化

認知を説明する神経生物学的に妥当なモデルは、パラメータがかなり複雑になります。そのようなモデルの構築と評価には、機械学習と大規模な脳・行動データセットが必要になります。従来、各研究室は、自分の専門分野の目標に焦点を当てて、独自のタスク、データセット、モデル、テストを開発してきました。しかし、このような取り組みを課題に合わせて拡大していくためには、3つの分野に関連するタスク、データ、モデル、テストを開発し、研究室間で共有する必要があります（図参照）。新しいコラボレーションの文化は、異なる研究室のコンポーネントを組み合わせることで、ビッグデータとビッグモデルを組み立てることになります。認知科学、計算論的神経科学、人工知能という 3 つの分野が一体となった成功の基準を満たすためには、従来の分野を超えた最適な役割分担が必要になるかもしれません。

タスク. 実験課題を設計することで、認知を定量的に調査可能な要素に切り分けます。タスクとは, 行動を制御するための環境のことです。タスクは、感覚入力（例：視覚刺激）と運動出力（例：ボタンを押す、ジョイスティックを操作する、より高次元の手足や全身の制御）を捉えるタスク「ワールド」のダイナミクスを定義します。タスクは、脳や行動のデータを取得し、AI モデルを開発する際に、明確な課題とモデルを比較するための定量的な性能ベンチマークを提供します。例えば、ImageNet タスク[148]は、コンピュータビジョンの大きな進歩をもたらしました。タスクは、データの取得とモデルの開発を推進するために、3つの分野すべてで容易に利用できるように設計・実装する必要があります（関連する開発には、OpenAI の [Gym](ttps://gym.openai.com/)、[Universe](https://universe.openai.com/)、DeepMind の Lab[149]があります)。役に立つタスクのスペクトルには、単純な刺激と反応を用いる古典的な心理物理学的タスクや、仮想現実でのインタラクションが含まれます。人間の心のあらゆる側面に関わるようになると、タスクは自然環境をシミュレートする必要があり、コンピュータゲームのようになっていくでしょう。これにより、特にタスクがインターネットを介して実行される場合には、大量の参加者と大きな行動データという付加的なメリットがもたらされる可能性があります[150]。

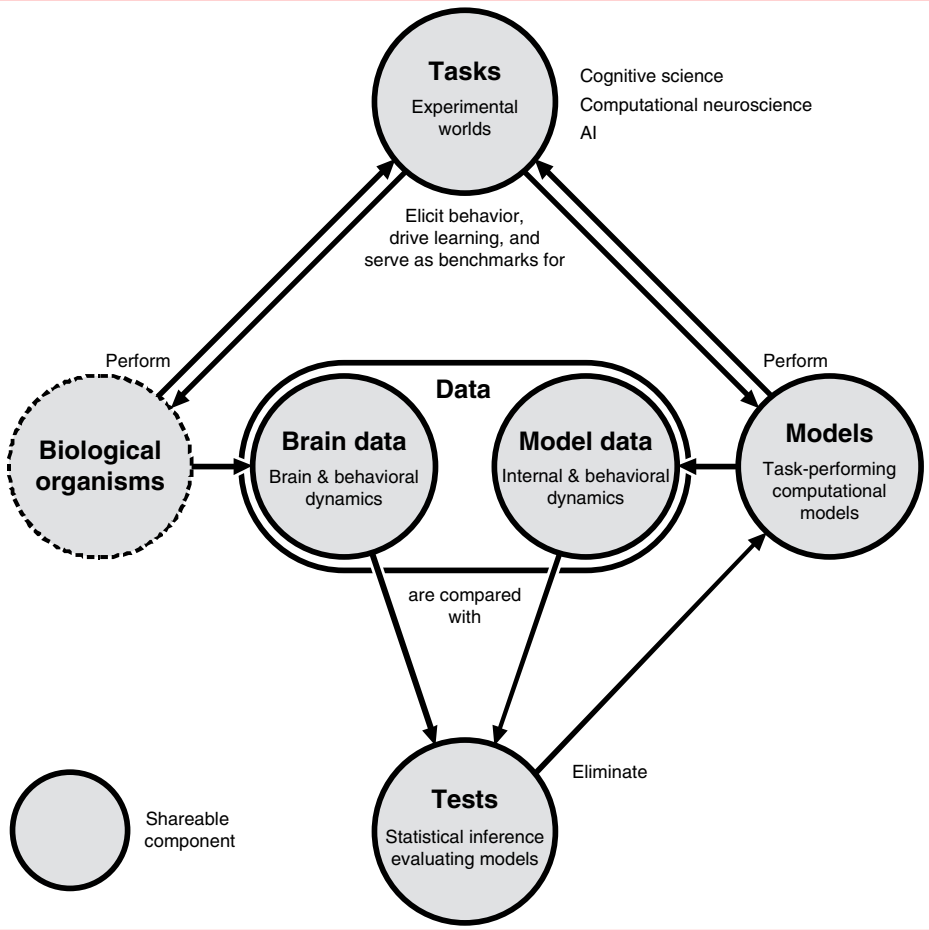
データ. タスクの実行中に得られた行動データは、全体的なパフォーマンスの推定値や、成功と失敗、反応時間と動作軌跡の詳

細なサインを提供します。脳活動の測定は、タスクパフォーマンスの基礎となるダイナミックな計算を特徴づける。解剖学的データは、脳の構造と接続性を複数のスケールで特徴づけることができます。脳の構造的データ、脳の機能的データ、行動的データのすべてが、計算モデルに制約を与えるために不可欠である。

モデル. タスクを実行する計算モデルは、実験的なタスクを実行するために、感覚入力を受け取り、運動出力を生成することができます。 AIスケールの神経生物学的に妥当なモデルは、オープンに共有され、そのタスクパフォーマンスや、モデル定義後に得られた新しいデータセットを含む様々な脳や行動のデータセットを説明する能力の観点からテストすることができる。モデルを定義した後に得られた新しいデータセットも含めて、様々な脳や行動のデータセットを説明する能力をテストします。最初は、多くのモデルが小さなタスクのサブセットに特化したものになるでしょう。最終的には、モデルはタスク間で一般化しなければならない。

テスト. あるモデルが特定のタスクにおける脳の情報処理をどの程度説明できるかを評価するためには、脳や行動のデータに基づいてモデルと脳を比較するテストが必要です。すべての脳は、その構造と機能において固有のものです。さらに、ある脳にとって、知覚、認知、行動のすべての行為は、時間的にユニークであり、まさにその脳を永久に変えてしまうため、繰り返すことができません。このような複雑さが、脳とモデルの比較を難しくしています。私たちは、関心のある要約統計と、モデルと脳の間の空間と時間における対応付けを、ある程度の抽象度で定義しなければなりません。モデル間の比較を行い、脳の理解にどれだけ近づいたかを判断するための適切なテストを開発することは、単に統計的推論の技術的課題ではありません。それは、理論的な神経科学の基本となる概念的な課題なのです。

研究室や分野間の相互作用は、敵対的な協力から利益を得ることができます[134]。現在の計算モデルでは認知の重要な側面を説明できないと感じている認知研究者は、これらの欠点を定量化する共有可能なタスクやテストを設計し、AI モデルの基準となる人間の行動データを提供するよう求められています。現在のモデルでは脳の情報処理を説明できないと感じている神経科学者は、タスク実行中に取得した脳の活動データや、脳とモデルの活動パターンを比較するテストを共有し、モデルの欠点を定量化することが求められています。成功の定義は複数あるでしょうが、それをモデルの質の定量的な尺度に変換することは不可欠であり、工学だけでなく認知計算神経科学の進歩にもつながるでしょう。



3　今後の展開

3.1　ボトムアップとトップダウン

脳は、ボトムアップの識別的計算とトップダウンの生成的計算をシームレスに統合して、知覚の推論や、モデルフリーとモデルベースの制御を行っています。脳科学も同様に、記述のレベルを統合し、ボトムアップとトップダウンの両方を進めていく必要があります。そうすることで、ニューロンのダイナミクスに基づいてタスクのパフォーマンスを説明し、脳がどのようにして心を生み出すのかをメカニズム的に説明することができます。

脳の計算を理解するために詳細な測定を行うというボトムアップの考え方は、最近の最も重要な資金調達の原動力となっています。欧州のヒューマン・ブレイン・プロジェクトや米国の BRAIN Initiative はいずれもボトムアップの考え方に基づいており、回路レベルに焦点を当てて脳のダイナミクスを測定・モデリングすることで、脳の計算を理解しようとしています。BRAIN Initiative は、神経細胞の活動を測定・操作する技術の向上を目指しています。 Human Brain Project は、神経科学のデータを生物学的に詳細な動的モデルに統合しようとするものです。いずれも、実験から理論へ、細胞レベルの記述から大規模な現象へと進んでいきます。

多数のニューロンを同時に測定し、その相互作用を回路レベルでモデル化することが不可欠となる。ボトムアップのビジョンは、科学の歴史に基づいています。例えば、顕微鏡や望遠鏡は、科学的なブレークスルーをもたらしました。しかし、より優れた観測結果によって理解が進むのは、常に先行する理論（観測されたプロセスの生成モデル）の中でのことです。例えば天文学では、コペルニクスの理論がガリレオの望遠鏡による観測結果を解釈する際の指針となりました。

脳を理解するためには、理論と実験を並行して開発し、ボトムアップのデータ駆動型アプローチを、説明すべき行動機能から始まるトップダウンの理論駆動型アプローチで補完する必要があります1[13,114]。これまでにないほど豊富な脳活動の測定と操作を行い、生物の行動適性に寄与する機能を果たすことができるかどうかという最初のテストに合格した脳-計算モデルを判断するために使用することで、理論的な洞察が得られます。このように、トップダウンのアプローチは、ボトムアップのアプローチを補完するものとして、脳の理解に欠かせないものとなっている（図3)。

3.2　マーのレベルを統合

Marr（1982）は、分析のレベルを 3 つに分けています。(1)計算理論、(2)表現とアルゴリズム、(3)神経生物学的実装である[115]。認知科学は、計算理論から始まり、認知を構成要素に分解し、表現とアルゴリズムをトップダウンで開発する。計算論的神経科学はボトムアップで進み、ニューロンの構成要素を、脳の全体的な機能の文脈で有用な構成要素と考えられる表現やアルゴリズムに構成する。計算神経科学はボトムアップで進められます。 AI は、単純な構成要素を組み合わせて複雑な知能を実現する表現やアルゴリズムを構築する。このように、3 つの学問分野は、脳と心のアルゴリズムと表現に収斂し、補完的な制約をもたらします[116]。

Marr のレベルは、脳を理解するための課題に役立つ指針となる。しかし、認知科学が脳を考慮する必要がないとか、計算論的神経科学が認知を考慮する必要がないということを示唆するものではありません（囲み記事 4)。 Marr は、コンピュータにインスピレーションを受けた。コンピュータは、人間の技術者が高レベルのアルゴリズムの記述に正確に適合するように設計する。これにより、技術者はアルゴリズムを設計する際に、回路を抽象化することができる。しかし、コンピュータサイエンスでも、アルゴリズムの一部は、並列処理能力などのハードウェアに依存します。しかし、脳はコンピュータとは異なり、この依存性をさらに強めている。脳は、進化と発達の産物であり、その過程では、ある抽象的な記述レベルで完璧に動作を把握できるシステムを生成するような制約はない。したがって、脳への実装を考えずに認知を理解することはできないし、逆に、認知機能を支える神経回路の文脈を無視して神経回路を理解することもできない。

分野を超えた課題の例として、初めてエスカレーターを見た子供の場合を考えてみましょう。エスカレーターを初めて見た子供は、人が斜め上に向かって昇っていくステップをすぐに認識します。エスカレーターを「動く階段」と考え、それに乗って、力を入れずに1階分持ち上げられることを想像するかもしれない。「エスカレーター」という言葉を覚える前に、たった一度の体験で機能を推察し、新しい概念を形成するかもしれない。

深層ニューラルネットワークモデルは、視覚体験の要素（人、段差、斜め上方向の動き、手すり）を迅速に認識することについて、生物学的に妥当な説明を提供する。計算効率の高いパターン認識コンポーネントを説明することができます[42]。しかし、

要素間の関係、物体の物理的な相互作用、人が上に行くという目的、エスカレーターの機能などを子供がどのように理解しているのか、また、体験を想像して瞬時に新しい概念を形成することができるのかについては、まだ説明できません。

ベイズノンパラメトリックモデルは、単一の経験からの深い推論や概念形成がいかにして可能かを説明する。脳の驚くべき統計的効率、すなわち、抽象的な事前知識を提供する生成モデルを構築することで、少ないデータから多くのことを推論する能力を説明できるかもしれません[98]。しかし，現在の推論アルゴリズムは大量の計算を必要とし，その結果，単一の視覚的経験から「エスカレーター」という新しい概念を形成するというような実世界の課題にはまだ対応できていない.

脳のアルゴリズムは、20 ワットの電力予算で、統計的効率と計算的効率を両立させており、ベイジアンやニューラルネットワーク型の現在の AI を凌駕しています。しかし、最近の AI や機械学習では、ベイジアン推論とニューラルネットワークモデルの交点を探り始めており、前者の統計的な強み（不確実性の表現、確率的推論、統計的効率）と後者の計算的な強み（表現的学習、普遍的な関数近似、計算効率）を組み合わせている[117-119]。

Marrの 3 つのレベルを統合するには、さまざまな専門知識を持つ研究者が緊密に協力する必要があります。神経科学、認知科学、AI スケールの計算モデリングを、一つの研究室が得意とすることは困難です。そのため、相補的な専門性を持つ研究室間でのコラボレーションが必要になります。従来の共同研究に加えて、分野間でコンポーネントを共有するオープンサイエンスの文化は、Marr のレベルを統合するのに役立ちます。共有できる要素としては、認知タスク、脳や行動のデータ、計算モデル、生物学的システムと比較してモデルを評価するテストなどがあります（囲み記事5）。

心と脳の研究は、特にエキサイティングな局面を迎えています。最近のコンピュータのハードウェアとソフトウェアの進歩により、AI スケールの心と脳のモデリングが可能になりました。認知科学、計算論的神経科学、AI が一体となれば、人間の認知を神経生物学的に妥当な計算モデルで説明できるようになるかもしれません。