# The brain produces mind by modeling

Richard M. Shiffrin[a,1], Danielle S. Bassett[b], Nikolaus Kriegeskorte[c], and Joshua B. Tenenbaum[d]

brain | mind | modeling

The connection of brain and mind has been a source of intense speculation at least since humanity became aware that the brain was the source of our behavior. Brain refers to the neurons, cells, and chemicals that govern activities of the organism. Mind is often considered consciously aware perceptions and thoughts. However, there is a gradient from unconscious to conscious, demonstrated by enormous amounts of research, such as the effects upon behavior of subliminal primes, so that mind is best considered to be the conscious and unconscious processes that act as an intermediate stage between the organism's biology and its behavior, or a translation from one to the other.

The National Academy of Sciences Colloquium "Brain Produces Mind by Modeling" was held May 1–3, 2019 at the Arnold and Mabel Beckman Center of the National Academy of Sciences in Irvine, CA. It was organized by Richard M. Shiffrin, Danielle S. Bassett, Nikolaus Kriegeskorte, and Joshua B. Tenenbaum. The theme of the colloquium and the foundation for the set of articles in this issue of PNAS is that the "mind" consists of a model formed by the "brain": This would be a model of the entire environment, including the self, the body, the physical environment, other agents, and the social environment. Furthermore, the model would be a best guess about the most likely state of this environment. It uses this model to learn, decide, attend, remember, perceive, predict, and produce action. This model develops as the brain matures, rapidly during infancy and more slowly later. It has structural components that remain stable over long times. It has labile elements that change at multiple time scales, adapting to the current environment and goals. The mind's formation through modeling of the world might be likened to the way scientists build models: through a combination of experiment (interaction with the world) and theory (thought) (1).

The existence of such a model is probably most evident from the many demonstrations that perception is a constructed model formed from sensory input. A simple example is the way we imagine seeing the entire forward field-of-view when we actually see clearly only a small foveally defined region (1). Such a construction is presumably formed on the basis of partial cues in regions that have poor acuity, from prior eye movements and from prior knowledge that tells us what is likely to be present (such as the likelihood that if we are in a room there will be walls and doors in peripheral regions of poor acuity). This model will of course be most useful for the purpose of maintaining a stable environment as the eyes, head, and body move, and for the purpose of guiding future eye movements (and bodily movements) as the need to do so arises (2).

There are innumerable visual "illusions" showing that perception is a construction of what is likely, but the idea that the brain produces mind by modeling is quite general, and applies to all aspects of cognition, including the social environment, such as the presence and motivations of other agents, and including our memories. A striking example of memory inference are demonstrations by Beth Loftus and colleagues (3) that we can form vivid and compelling memories of events that never happened. That high-level social inference is a model is postulated in the "theory of mind," raised by Premack and Woodruff in 1978 (4) about nonhuman primates. This theme occurs in what has been called the "computational theory of mind." It is found in a variety of attempts to model learning and behavior based on processes such as

[a]Psychological and Brain Sciences Department, Indiana University, Bloomington, IN 47405; [b]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104; [c]Zuckerman Mind Behavior Institute, Columbia University, New York, NY 10027; and [d]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139-4307

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Brain Produces Mind by Modeling," held May 1–3, 2019, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. NAS colloquia began in 1991 and have been published in PNAS since 1995. From February 2001 through May 2019, colloquia were supported by a generous gift from The Dame Jillian and Dr. Arthur M. Sackler Foundation for the Arts, Sciences, & Humanities, in memory of Dame Sackler's husband, Arthur M. Sackler. The complete program and video recordings of most presentations are available on the NAS website at http://www.nasonline.org/brain-produces-mind-by.

Author contributions: R.M.S., D.S.B., N.K., and J.B.T. wrote the paper.

The authors declare no competing interest.



[1]To whom correspondence may be addressed. Email: shiffrin@indiana.edu.

First published November 23, 2020.

prediction error and surprise (5–7), represented in neural net modeling in cognitive science (6, 7), and in artificial intelligence (AI) and machine learning, as represented by error-driven learning and reinforcement learning (7) and implemented by inference algorithms using Markov chain Monte Carlo and belief propagation (8). Thus, a number of efforts have tried to bridge the gap from brain to mind by building models capable of feats of cognition whose component computations might plausibly be implemented in neural circuitry. The theme of brain as model builder is also central to the Bayesian approach in cognitive science, where judgments and decisions are explained by probabilistic inference, combining prior knowledge and current evidence (9). It is worth noting that the theme of "brain as scientist" could be extended to "mind as scientist," surely true to the extent that brain produces mind. Although a long history of research demonstrates the occasional irrationality of human decision making (10), the colloquium and the articles in this issue aim to present a coherent view of the brain as a model builder, the model being used to maximize survival and utility in a complex world.

The 13 articles in this special issue of PNAS that resulted from the Colloquium are next briefly described. Although these articles were inspired by the conference theme, most represent current state-of-the-art research by the authors, so although there is a connection to the theme, that connection may not always be immediately obvious.

Kelsey R. Allen, Kevin A. Smith, and Joshua B. Tenenbaum have contributed an article titled "Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning" (11). The brain's model of the environment includes a model of physical dynamics and supports creative physical reasoning, a quintessentially human ability. The authors introduce a novel task to study this reasoning in both humans and machines. They provide a computational model that uses structured priors and physical simulation to perform in a human-like way.

Zhengwei Wu, Minhae Kwon, Saurabh Daptardar, Paul Schrater, and Xaq Pitkow contribute an article titled "Rational thoughts in neural codes" (12). They offer a way to explain an animal's behavior as rational—optimal for a mistaken model of the world—and show how to interpret brain activity as encoding and transforming its rational thoughts.

Douglas A. Ruff, Cheng Xue, Lily E. Kramer, Faisal Baqai, and Marlene R. Cohen contribute an article titled "Low rank mechanisms underlying flexible visual representations" (13). The authors study how the brain models, represents, and interprets a wide range of sensory and cognitive processes, and how that model incorporates neural variability. They find that several sensory and cognitive processes have similarly low-dimensional effects on the variability of populations of visual neurons.

Tal Golan, Prashant C. Raju, and Nikolaus Kriegeskorte contribute an article titled "Controversial stimuli: Pitting neural networks against each other as models of human cognition" (14). They introduce controversial stimuli, a method to generate visual images about which neural network models disagree, thereby revealing their distinct inductive biases. They show that models employing generative inference perform better than purely discriminative models at predicting the categories of objects humans recognize in controversial stimuli. This is consistent with the idea that human recognition is an inference process that recovers the things in the world as they would have to be to explain our sensations, echoing observations by Helmholtz in 1867 in his "Treatise on physiological optics: Concerning the perceptions in general" (15).

Neal W. Morton, Margaret L. Schlichting, and Alison R. Preston contribute an article titled: "Representations of common event structure in medial temporal lobe and frontoparietal cortex support efficient inference" (16). Events often share abstract structure; for example, restaurant visits involve ordering, eating, and paying the bill. They argue that representational geometry in the medial temporal lobe and parietal cortex reflects abstract structure and supports reasoning about events. The authors propose that abstract representations act as a model to guide retrieval of specific events from a cognitive map.

David Stawarczyk, Christopher N. Wahlheim, Joset A. Etzel, Abraham Z. Snyder, and Jeffrey M. Zacks contribute an article titled "Aging and the encoding of changes in events: The role of neural activity pattern reinstatement" (17). Using pattern-based fMRI, the authors investigated how memory representations act as a model to guide expectations about upcoming events in young and older adults. The authors found that memory representation reinstatement affected new encoding in both groups, and that reinstatement was associated with better memory updating when events changed.

Emelie L. Josephs and Talia Konkle contribute an article titled "Large-scale dissociations between views of objects, scenes, and reachable-scale environments in visual cortex" (18). The human brain must accurately represent and model views of the reachable world. The authors identify a distinct topography of visual responses to reachable views compared to scene and object views, including the existence of three novel brain regions with preference for reachable views over both scenes and objects. This suggests that the brain's modeling of rich near-scale views is computationally dissociable from that of navigable-scale scenes and singleton objects.

Stephen Sebastian, Eric S. Seemiller, and Wilson S. Geisler have contributed an article titled "Local reliability weighting explains identification of partially masked objects in natural images" (19). Our brains have learned, through evolution and experience, the statistical properties of our natural environments and exploit this knowledge when performing perceptual tasks. The authors identify a new statistical property of natural images, the "partial-masking factor," that is crucial for object identification. They show it is exploited by the human visual system in a near-optimal way, consistent with Bayesian inference.

Chaitanya K. Ryali, Stanny Goffin, Piotr Winkielman, and Angela J. Yu contribute an article titled "From likely to likable: The role of statistical typicality in human social assessment of faces" (20). The paper provides a novel theoretical argument and experimental evidence for how the human brain represents statistical typicality of a face, the way that typicality contributes to positive social impressions from faces, such as attractiveness and trustworthiness, and how informational needs and attentional prioritization modulate internal statistical representation, and consequently, key social impressions.

Maria K. Eckstein and Anne G. E. Collins contribute the article: "Computational evidence for hierarchically structured reinforcement learning in humans" (21). The authors present the case for one of the most firmly held beliefs in cognitive science, that the model of mind the brain forms is hierarchically structured. The authors provide computational evidence for this claim, showing that hierarchical reinforcement learning provides a biologically realistic and computationally simple account of human learning and generalization.

Mathilee Kunda contributes an article titled "AI, visual imagery, and a case study on the challenges posed by human intelligence

tests" (22). To solve a visuo-spatial intelligence test, the brain must at minimum form a visuo-spatial model, if not a complete model of the agent's environment. How the brain might form such a model is explored with AI, in the form of computational models based on human visual imagery that are aimed to approach human capabilities.

Giwon Bahg, Daniel G. Evans, Matthew Galdo, and Brandon M. Turner contribute an article entitled "Gaussian process linking functions for mind, brain, and behavior" (23). This work proposes a statistical approach to capturing an observer's state of mind based on the dynamics simultaneously present in neural and behavioral data.

Christopher W. Lynn and Danielle S. Bassett contribute an article titled "How humans learn and represent networks" (24). A large part of the mind's model of the environment is its various networks, since most information in the environment is in the form of networks, social and otherwise. These authors describe how network knowledge might be learned.


1 A. H. S. Chan, A. J. Courtney, Foveal acuity, peripheral acuity, and search performance: A review. *Int. J. Ind. Ergon.* **18**, 113–119 (1996).
2 J. Najemnik, W. S. Geisler, Eye movement statistics in humans are consistent with an optimal search strategy. *J. Vis.* **8**, 1–14 (2008).
3 D. M. Bernstein, C. Laney, E. K. Morris, E. F. Loftus, False beliefs about fattening foods can have healthy consequences. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 13724–13731 (2005).
4 D. Premark, G. Woodruff, Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* **4**, 515–526 (1978).
5 J. Gläscher, N. Daw, P. Dayan, J. P. O'Doherty, States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).
6 T. T. Rogers, J. L. McClelland, Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cogn. Sci.* **38**, 1024–1077 (2014).
7 R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998).
8 Y. Weiss, W. T. Freeman, Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Comput.* **13**, 2173–2200 (2001).
9 T. L. Griffiths, C. Kemp, J. B. Tenenbaum, "Bayesian models of cognition" in *The Cambridge Handbook of Computational Cognitive Modeling*, R. Sun, Ed. (Cambridge University Press, 2008).
10 Z. Wang, T. Solloway, R. M. Shiffrin, J. R. Busemeyer, Context effects produced by question orders reveal quantum nature of human judgments. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 9431–9436 (2014).
11 K. R. Allen, K. A. Smith, J. B. Tenenbaum, Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29302–29310 (2020).
12 Z. Wu, M. Kwon, S. Daptardar, P. Schrater, X. Pitkow, Rational thoughts in neural codes. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29311–29320 (2020).
13 D. A. Ruff, C. Xue, L. E. Kramer, F. Baqai, M. R. Cohen, Low rank mechanisms underlying flexible visual representations. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29321–29329 (2020).
14 T. Golan, P. C. Raju, N. Kriegeskorte, Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29330–29337 (2020).
15 H. Helmholtz, "Handbuch der physiologischen Optik" in *Allgemeine Encyklopddie der Physik*, G. Karsten, Ed. (Voss, Leipzig, 1867), vol. 9.
16 N. W. Morton, M. L. Schlichting, A. R. Preston, Representations of common event structure in medial temporal lobe and frontoparietal cortex support efficient inference. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29338–29345 (2020).
17 D. Stawarczyk, C. N. Wahlheim, J. A. Etzel, A. Z. Snyder, J. M. Zacks, Aging and the encoding of changes in events: The role of neural activity pattern reinstatement. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29346–29353 (2020).
18 E. L. Josephs, T. Konkle, Large-scale dissociations between views of objects, scenes, and reachable-scale environments in visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29354–29362 (2020).
19 S. Sebastian, E. S. Seemiller, W. S. Geisler, Local reliability weighting explains identification of partially masked objects in natural images. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29363–29370 (2020).
20 C. K. Ryali, S. Goffin, P. Winkielman, A. J. Yu, From likely to likable: The role of statistical typicality in human social assessment of faces. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29371–29380 (2020).
21 M. K. Eckstein, A. G. E. Collins, Computational evidence for hierarchically structured reinforcement learning in humans. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29381–29389 (2020).
22 M. Kunda, AI, visual imagery, and a case study on the challenges posed by human intelligence tests. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29390–29397 (2020).
23 G. Bahg, D. G. Evans, M. Galdo, B. M. Turner, Gaussian process linking functions for mind, brain, and behavior. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29398–29406 (2020).
24 C. W. Lynn, D. S. Bassett, How humans learn and represent networks. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29407–29415 (2020).