

# Bayes in the Age of Intelligent Machines

Thomas L. Griffiths<sup>1,2</sup> , Jian-Qiao Zhu<sup>1,2</sup>, Erin Grant<sup>3</sup>,  
and R. Thomas McCoy<sup>4</sup>

<sup>1</sup>Department of Psychology, Princeton University, <sup>2</sup>Department of Computer Science, Princeton University,

<sup>3</sup>Gatsby Computational Neuroscience Unit, University College London, and <sup>4</sup>Department of Linguistics,  
Yale University

Current Directions in Psychological  
 Science  
 2024, Vol. 33(5) 283–291  
 © The Author(s) 2024  
 Article reuse guidelines:  
[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)  
 DOI: 10.1177/09637214241262329  
[www.psychologicalscience.org/CDPS](http://www.psychologicalscience.org/CDPS)



## Abstract

The success of methods based on artificial neural networks in creating intelligent machines seems like it might pose a challenge to explanations of human cognition in terms of Bayesian inference. We argue that this is not the case and that these systems in fact offer new opportunities for Bayesian modeling. Specifically, we argue that artificial neural networks and Bayesian models of cognition lie at different levels of analysis and are complementary modeling approaches, together offering a way to understand human cognition that spans these levels. We also argue that the same perspective can be applied to intelligent machines, in which a Bayesian approach may be uniquely valuable in understanding the behavior of large, opaque artificial neural networks that are trained on proprietary data.

## Keywords

Bayesian modeling, computational modeling, artificial intelligence

In the 18th century, Thomas Bayes had a radical idea: using probabilities to represent the degrees to which we believe hypotheses are true (Bayes, 1763/1958). He did so in the context of a gambling game: Having seen some number of wins and losses, how likely are you to win? The idea of using probability theory to update our degrees of belief on the basis of data underlies what we now call Bayes' rule (see Fig. 1). Bayes would presumably have assigned low probability to his work becoming the foundation, more than 2 centuries later, for Bayesian models of cognition, which explain human behavior in terms of rational belief updating (e.g., Griffiths et al., 2010).

Bayesian models of cognition explain inductive inference—the process of going from limited data to an uncertain conclusion, such as inferring the meaning of a new word on the basis of hearing that word in conversation. In Bayesian models, such inferences are framed as the result of combining data (e.g., the context in which you heard the new word) with our existing expectations about the world (e.g., expectations about what sorts of meanings a word could have). Those expectations are expressed in a “prior distribution” over hypotheses, with more plausible hypotheses having higher prior probability. This captures the “inductive biases” of a learner—those factors other than the data that influence the hypothesis the learner selects (Mitchell, 1997). Prior distributions can be defined over

complex and expressive hypotheses, including grammars, causal structures, logical formulas, and programs, providing a way to characterize inductive biases using structured symbolic models while still supporting the ability to learn (e.g., Goodman et al., 2011; Griffiths & Tenenbaum, 2009; Rule et al., 2020). For example, a model of language learning could postulate that learners consider a set of possible grammars, with different prior probabilities assigned to each of those grammars (Yang & Piantadosi, 2022).

Part of the appeal of the Bayesian approach is explaining behavior via the rational solution to an abstract problem. If we accept that degrees of belief should be expressed as probabilities, then Bayes' rule solves the problem of inductive inference. This creates the opportunity to discover connections to other disciplines: Statisticians and computer scientists also want to create systems that make inferences from limited data. Bayesian models of cognition benefited from these connections because discoveries from statistical machine learning informed accounts of human cognition (e.g., Sanborn et al., 2010). However, the last decade has seen a significant change in the landscape

## Corresponding Author:

Thomas L. Griffiths, Departments of Psychology and Computer  
 Science, Princeton University  
 Email: [tomg@princeton.edu](mailto:tomg@princeton.edu)

$$P(\text{hypothesis} \mid \text{data}) = \frac{P(\text{data} \mid \text{hypothesis}) P(\text{hypothesis})}{P(\text{data})}$$

**Fig. 1.** Bayes' rule. The formula indicates how a rational agent should update their beliefs, providing a mathematical specification for optimal inductive inference. It establishes a relationship between the prior belief in a hypothesis,  $P(\text{hypothesis})$ , and the updated posterior belief,  $P(\text{hypothesis} \mid \text{data})$ , after incorporating evidence from data. The term  $P(\text{data} \mid \text{hypothesis})$  represents the probability of observing the given data if the hypothesis is true, whereas  $P(\text{data})$  acts as a normalizing constant that guarantees the resulting probabilities sum to 1.

of machine learning. Major breakthroughs have been the result not of more sophisticated Bayesian methods but of increasingly large artificial neural networks that are trained on increasingly large amounts of data (LeCun et al., 2015). Does the success of this approach in creating intelligent machines undermine the importance of Bayes' rule as a tool for understanding human cognition?<sup>1</sup>

In this article, we consider the prospects for Bayes in the age of intelligent machines. We first argue that the success of large artificial neural networks—"deep learning" (LeCun et al., 2015)—does not pose a challenge for Bayesian models of cognition and is actually complementary. We then argue that Bayesian models have an important new application: understanding the behavior of those intelligent machines. Deep learning has been successful in creating systems that can solve challenging problems, but the resulting systems are opaque and difficult to analyze. We suggest that the methods psychologists have developed for understanding an equally opaque and difficult-to-analyze system—human beings—can be adapted to make sense of large artificial neural networks and that Bayesian models in particular have insights to offer. The key to both of these arguments is the idea of levels of analysis, which we explore in more detail in the next section.

## Levels of Analysis

David Marr famously argued that information processing systems can be understood at multiple levels of analysis (Marr, 1982). To borrow an analogy suggested by Marr, a biologist interested in understanding bird flight could pursue that goal in different ways. They could focus on the physical mechanisms underlying flight, asking how muscles, bones, and feathers translate into lift; on the procedures that birds use to change their wing position to take off, glide, and land; or on the abstract principles of aerodynamics that determine why bird wings have a particular shape. When we study information processing systems—including humans and

machines—we have the same kind of choice about the level at which we analyze those systems.

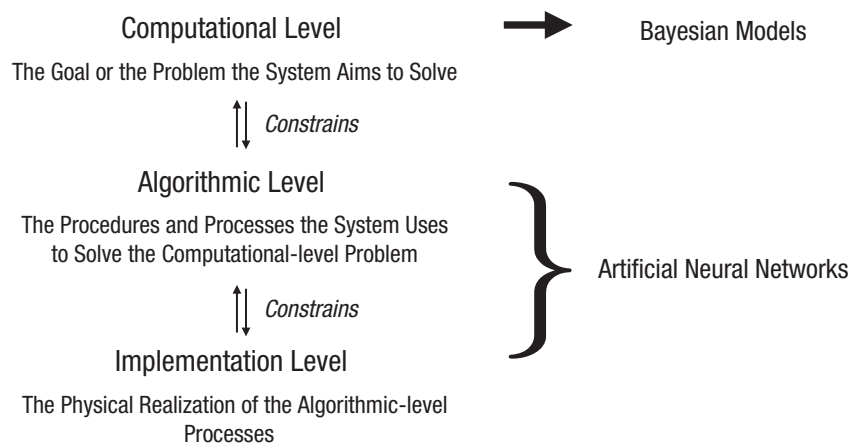
Marr laid out three different levels at which we can analyze an information processing system (see Fig. 2). The most abstract is the *computational* level, at which we consider the problem that the system is solving and what an ideal solution to that problem looks like.<sup>2</sup> If the underlying problem involves learning or inductive inference, then an ideal solution to that problem comes from Bayes' rule—Bayesian models of cognition are defined at this level of analysis. Next is the *algorithmic* level, at which we consider what algorithm might (perhaps approximately) solve this problem and what representations it operates over. Finally, there is the *implementation* level, at which we ask how that representation and algorithm can be realized physically.<sup>3</sup>

Marr also came out strongly in favor of pursuing questions at one level of analysis in particular—the computational level. He wrote that

trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: It just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense. (Marr, 1982, p. 27)

Marr's argument inspired a generation of researchers pursuing computational-level analyses of cognition (Anderson, 1990; Oaksford & Chater, 1994; Shepard, 1987; Tenenbaum & Griffiths, 2001). However, understanding human cognition is ultimately going to require answers at all three levels of analysis.

Marr's levels of analysis reveal that there are different kinds of questions that we can ask about information processing systems, each with a corresponding kind of answer, and that those answers are not necessarily in conflict with one another. For example, cognitive scientists studying human memory could offer theories at each level—one an optimal solution to a computational



**Fig. 2.** Marr's levels of analysis. Marr (1982) provided a framework for understanding information processing systems such as the human brain or AI systems. Different kinds of computational models engage with these different levels—Bayesian models are typically defined at the computational level, whereas artificial neural networks explore hypotheses at the algorithmic and implementation levels.

problem, one a cognitive process, and one a neural circuit—and those theories could all be correct. The key is that the theories need to be compatible: The processes at the algorithmic level must result in a reasonable approximation to the solution at the computational level, and neurons at the implementation level must, in turn, execute something like that algorithm.

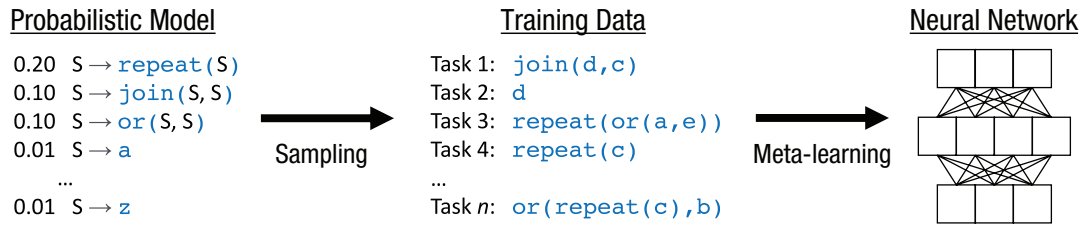
### Bayes and Deep Learning Are Complementary Approaches

Having introduced the idea of levels of analysis, we can now make our first argument: The success of deep learning is not a challenge to Bayesian models of cognition because these two approaches address different levels of analysis and are compatible. The first of these claims is relatively straightforward: As noted above, Bayesian models of cognition are explicitly defined at the computational level. By contrast, accounts of human cognition based on artificial neural networks typically situate themselves at the algorithm or implementation levels, focusing on cognitive or neural processes rather than abstract problems and their ideal solution (e.g., McClelland et al., 2010). The key issue is thus whether these approaches are compatible.

There are both theoretical arguments and empirical results supporting the view that they are compatible. As noted above, Bayes' rule is an optimal solution to problems of inductive inference assuming that the world is well described by a particular prior distribution over hypotheses. Artificial neural networks are trained to minimize a loss function (also known as an error function or an objective function) that measures how well it is performing a particular task. Those loss functions have natural probabilistic interpretations that can

be used to relate them to Bayesian inference.<sup>4</sup> Artificial neural networks thus have a direct interpretation as a kind of probabilistic model and should be seeking a hypothesis (i.e., a set of connection weights) that assigns high probability to the observed data. We thus need to show that these models capture the impact of the prior distribution on selecting that hypothesis in a way that is consistent with Bayesian inference.

Classic theoretical results show that the algorithms used to train artificial neural networks—decreasing the weights after each step of training (MacKay, 1995) or deliberately taking only a few passes through the data (Bishop, 1995)—are consistent with imposing a Gaussian prior on the weights of the network. Other analyses show how specific neural network architectures can be used to implement Bayesian inference for arbitrary prior distributions (Shi & Griffiths, 2009). More recently, researchers have developed methods for performing Bayesian inference using neural networks by explicitly training networks to approximate the output of Bayes' rule for a specific prior distribution (Dasgupta & Gershman, 2021). Finally, empirical analyses of deep learning models show that they seem to internalize information that can be used to perfectly reconstruct relevant Bayesian posterior distributions (Mikulik et al., 2020). Even if the strategies that deep learning systems are using are not consistent with Bayesian inference with any particular prior, a Bayesian ideal is at least a starting point in making sense of these systems (e.g., Raventós et al., 2023). Systematic deviations from that Bayesian ideal would suggest the behavior could be explained as approximate Bayesian inference under resource constraints, an idea that has also been used in understanding human cognition (Griffiths et al., 2015).



**Fig. 3.** Distilling a Bayesian model’s prior into a neural network. We first define a prior using a probabilistic model (left). We then sample many tasks from that prior (middle). Here, each “task” is a formal rule defining a set of strings; for example, *repeat(c)* defines the set containing *c*, *cc*, *ccc*, and so on. Finally, metalearning is used to create a neural network that is trained to perform those tasks (right), giving it a prior that approximates the one we started with. Metalearning is a process in which a learner encounters many different tasks and leverages the commonalities across those tasks to gain inductive biases that enable it to learn new tasks more easily.

But what about the grammars, causal structures, logical formulas, and programs that play a prominent role in Bayesian models of cognition? There is nothing like those built into artificial neural networks, which represent the world with continuous weights and activations. This is problematic if the structured representations used in defining Bayesian models are interpreted as a claim about the algorithmic and implementation levels, which are the levels that artificial neural networks are designed to model. However, it is not problematic if we view those representations as being primarily useful in providing a way to specify human-like inductive biases. In this case, the question becomes whether artificial neural networks can manifest inductive biases consistent with those specified using such structured representations.

Grant et al. (2018) suggested that this is the case: They showed that meta-learning—a procedure in which the initial weights of a neural network are adapted to make it easier for that network to perform a variety of tasks—can be interpreted as learning an appropriate Bayesian prior distribution for those tasks. We have built on that suggestion to develop a method for “distilling” an explicit prior distribution from a Bayesian model into a neural network (McCoy & Griffiths, 2023; see Fig. 3). This method generates a set of tasks by sampling from that prior distribution and then uses metalearning to create a neural network that is easily adapted to perform those tasks.

McCoy and Griffiths (2023) showed that this approach can distill an abstract prior distribution over formal languages—itsself defined by a grammar—into a set of initial weights for a recurrent neural network. The resulting neural network can learn new formal languages from the same amount of data as a Bayesian model that uses a prior distribution defined over symbolic grammars (Yang & Piantadosi, 2022). An analysis of the behavior of this network shows that the learned initial weights induce a bias toward recursive structures—exactly what is required for modeling both formal and natural languages. We anticipate that a similar

approach will be effective for other structured prior distributions, providing the last piece of evidence that deep learning and Bayesian models are complementary approaches to understanding the mind.

Although we have focused on connecting computational-level models to neural networks, similar opportunities hold for algorithmic-level models derived from the Bayesian approach. One popular strategy for creating algorithmic-level models of inductive inference has been to draw on the idea that people might approximate the relevant probability distributions by sampling (Griffiths et al., 2012; Sanborn et al., 2010). These sampling algorithms can in turn be implemented by neural networks (e.g., Shi & Griffiths, 2009), providing further avenues for developing complementary models.

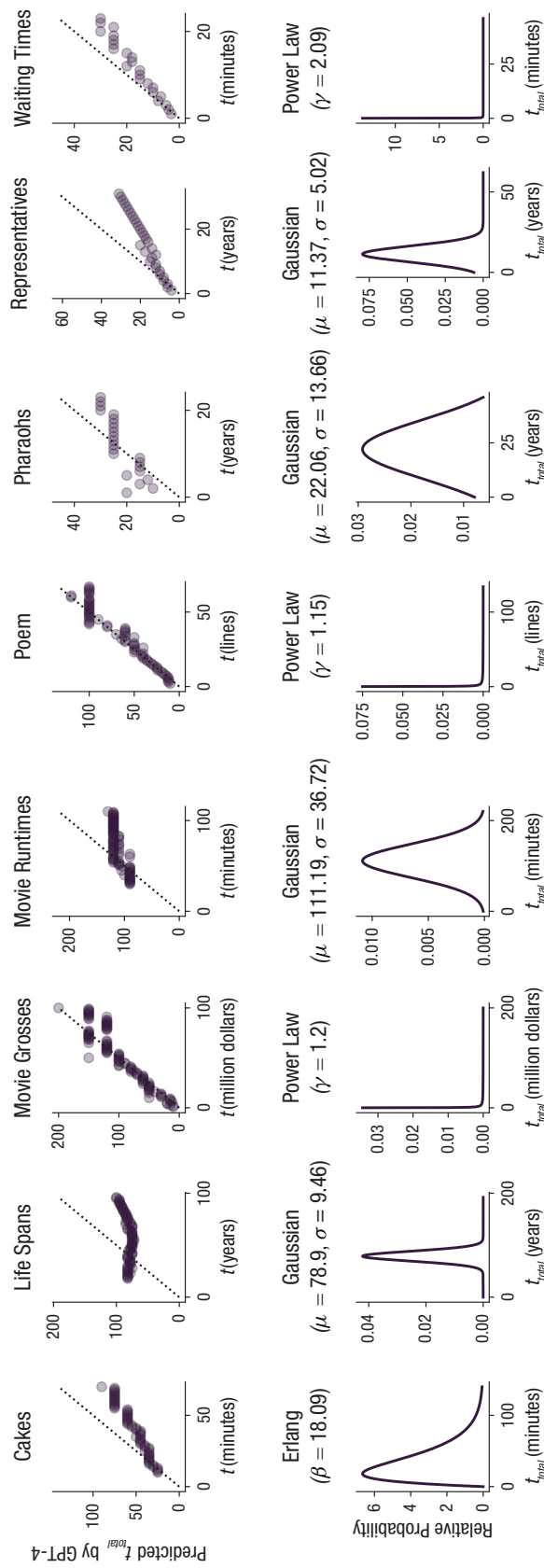
## From Modeling People to Modeling Machines

The claim that Bayesian models of cognition and deep learning are complementary has a broader implication: that we can expand the scope of Bayesian modeling from humans to machines. Deep learning has created systems that can solve challenging problems, but it has a number of limitations. Deep neural networks are opaque and hard to interpret, particularly when their internal workings are withheld by the companies that create them. This leaves computer scientists in the unfamiliar territory of trying to make sense of complex information processing systems via their behavior. This, of course, is a problem that psychologists are intimately familiar with and a place where Bayesian models of cognition might be uniquely helpful.<sup>5</sup>

We can apply Marr’s levels of analysis to AI systems. This idea is novel in the context of machine learning, in which a practitioner might just think about choosing a method—Bayes or deep learning—to use to solve an engineering problem. But the compatibility of these approaches means that even if the practitioner chooses







**Fig. 5.** GPT-4's forecasts for different everyday events. Baking times for cakes; life spans; movie grosses; movie run times; poem lengths; reigns of Egyptian Pharaohs; terms of members of the U.S. House of Representatives; and waiting times for calling a box office are shown from left to right, respectively. The top row of plots illustrates GPT-4's predictions of total duration or extent,  $t_{\text{total}}$ , for each sample event duration  $t$  (dots). Dashed lines represent predictions using a fixed noninformative prior, which predicts  $t_{\text{total}}$  should simply be twice  $t$ . The bottom row presents the recovered prior distributions of  $t_{\text{total}}$  for each event based on GPT-4's predictions using a Bayesian model presented in Griffiths and Tenenbaum (2006).  $\beta$  represents the scale of an Erlang distribution,  $\mu$  is the mean and  $\sigma$  is the standard deviation in a Gaussian distribution, and  $\gamma$  indicates power in a power-law distribution.

Ribeiro et al., 2016) and has begun to explore how Bayesian models can augment these approaches by providing insight into levels of uncertainty (e.g., Wang et al., 2023). The approach we have outlined here adds to this the idea that we might be able to quantify the *inductive biases* of AI systems by explicitly modeling them as making Bayesian inferences, and that by doing so we are able to make direct comparisons between the inductive biases of machines and those of humans.

## Conclusion

The success of artificial neural networks in machine learning would seem to pose a challenge to Bayesian models of cognition. Instead, we argue, it presents an opportunity. First, it provides a way to begin to explain human behavior at multiple levels of analysis, with Bayesian models at the computational level and neural networks at the algorithmic and implementation levels. Second, the artificial neural networks used in machine learning are opaque, complex, and difficult to interpret—just like humans. Bayesian models thus provide a new tool for exploring the inductive biases of machines. To echo Marr, it seems unlikely that we will understand cognition by studying only artificial neurons.

## Recommended Reading

- Anderson, J. R. (1990). (See References). Presents a detailed discussion of the idea of levels of analysis and rational models of cognition.
- Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). (See References). Discusses the motivation behind Bayesian models of cognition in contrast to connectionist approaches.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). (See References). Presents a brief overview of the ideas behind deep learning.
- McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). (See References). Provides a response to Griffiths et al. (2015) and a critique from the connectionist perspective.
- Mitchell, M. (2023). (See References). Highlights the challenges of evaluating the capacities of AI systems.

## Transparency

*Action Editor:* Robert L. Goldstone

*Editor:* Robert L. Goldstone

*Declaration of Conflicting Interests*

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

*Funding*

This work was supported by the NOMIS Foundation and National Science Foundation Social, Behavioral and Economic Sciences Postdoctoral Research Fellowship Grant 2204152.

## ORCID iD

Thomas L. Griffiths  <https://orcid.org/0000-0002-5138-7255>

## Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/09637214241262329>

## Notes

1. It is worth noting that there are still settings in which Bayesian methods are favored in machine learning, particularly settings in which there are limited data, interpretability matters, and quantifying uncertainty is important, such as in science or medicine (e.g., Alaa & Van Der Schaar, 2017; Padilla et al., 2021; Wang et al., 2023). Bayesian methods also underlie some of the most successful approaches to image generation (e.g., Kingma & Welling, 2013; Song et al., 2021).
2. Biological systems, as opposed to technological ones, are unlikely to be characterized as a whole in terms of solving a single problem we can easily identify. For this reason, computational-level models typically focus on isolated aspects of cognition for which it is comparatively straightforward to postulate a specific problem and identify its solution (for more details, see Anderson, 1990).
3. Cognitive scientists have considered a variety of other ways of formulating these levels of analysis, including adding levels to capture finer grained distinctions (for a review, see Anderson, 1990, Chapter 1). We use the levels introduced by Marr to make our point here, but we imagine that similarly fine-grained distinctions will become apparent as we deepen our understanding of intelligent machines.
4. For example, minimizing the cross-entropy loss corresponds to maximizing the probability of discrete data, and minimizing the squared-error loss corresponds to maximizing the probability of continuous data under a Gaussian distribution (for further details, see MacKay, 1995).
5. Although our focus here is on the role of Bayesian models of cognition in this new setting, it is worth noting that the methods of cognitive modeling in general are potentially useful in understanding modern AI systems. In particular, connectionist modelers working with smaller scale neural networks developed an effective toolbox for analyzing the information contained in the weights of those networks (e.g., Rodriguez et al., 1999; Rogers & McClelland, 2004), and some of those tools might be productively transferred to the analysis of large-scale neural networks.

## References

- Alaa, A. M., & Van Der Schaar, M. (2017). Bayesian inference of individualized treatment effects using multi-task Gaussian processes. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30 (NIPS 2017)*. Curran Associates, Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/6a508a60aa3bf9510ea6acb021c94b48-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/6a508a60aa3bf9510ea6acb021c94b48-Paper.pdf)
- Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum & Associates.

- Bayes, T. (1958). Studies in the history of probability and statistics: IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances. *Biometrika*, *45*, 296–315. (Original work published 1763)
- Bishop, C. (1995). Regularization and complexity control in feed-forward networks. In *Proceedings of the International Conference on Artificial Neural Networks* (pp. 141–148). Institute of Electrical and Electronics Engineers.
- Dasgupta, I., & Gershman, S. J. (2021). Memory as a computational resource. *Trends in Cognitive Sciences*, *25*(3), 240–251.
- Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*, 110–119.
- Grant, E., Finn, C., Levine, S., Darrell, T., & Griffiths, T. L. (2018, April 30–May 3). *Recasting gradient-based meta-learning as hierarchical Bayes* [Conference session]. 6th International Conference on Learning Representations, Vancouver, BC, Canada.
- Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.
- Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229.
- Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 767–773.
- Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661–716.
- Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*, 263–268.
- Kingma, D. P., & Welling, M. (2013, May 2–4). *Auto-encoding variational Bayes* [Conference session]. 1st International Conference on Learning Representations, Scottsdale, AZ, United States.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.
- Li, M. Y., Grant, E., & Griffiths, T. L. (2023). Gaussian process surrogate models for neural networks. In R. J. Evans & P. Shpitser (Eds.), *UAI'23: Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence* (pp. 1241–1252). Association for Computing Machinery.
- Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215.
- MacKay, D. (1995). Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, *6*, 469–505.
- Marr, D. (1982). The philosophy and the approach. In *Vision* (pp. 8–37). W. H. Freeman and Company.
- McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to understanding cognition. *Trends in Cognitive Sciences*, *14*, 348–356.
- McCoy, R. T., & Griffiths, T. L. (2023). *Modeling rapid language learning by distilling Bayesian priors into artificial neural networks*. arXiv. <https://doi.org/10.48550/arXiv.2305.14701>
- McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). *Embers of autoregression: Understanding large language models through the problem they are trained to solve*. arXiv. <https://doi.org/10.48550/arXiv.2309.13638>
- Mikulik, V., Delétang, G., McGrath, T., Genewein, T., Martic, M., Legg, S., & Ortega, P. (2020). Meta-trained agents implement Bayes-optimal agents. In H. Larochelle et al. (Eds.), *Advances in neural information processing systems 33 (NIPS 2020)*. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2020/file/d902c3ce47124c66ce615d5ad9ba304f-Paper.pdf>
- Mitchell, M. (2023). How do we know how smart AI systems are? *Science*, *381*(6654), Article eadj5957. <https://doi.org/10.1126/science.adj5957>
- Mitchell, T. M. (1997). *Machine learning*. McGraw Hill.
- Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.
- Padilla, L. E., Tellez, L. O., Escamilla, L. A., & Vazquez, J. A. (2021). Cosmological parameter inference with Bayesian statistics. *Universe*, *7*(7), Article 213. <https://doi.org/10.3390/universe7070213>
- Raventós, A., Paul, M., Chen, F., & Ganguli, S. (2023). Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In A. Oh et al. (Eds.), *Advances in neural information processing systems 36 (NIPS 2023)*. Curran Associates, Inc. [https://papers.nips.cc/paper\\_files/paper/2023/file/2e10b2c2e1a4f8083c37dfe269873f8-Paper-Conference.pdf](https://papers.nips.cc/paper_files/paper/2023/file/2e10b2c2e1a4f8083c37dfe269873f8-Paper-Conference.pdf)
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. In *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery.
- Rodriguez, P., Wiles, J., & Elman, J. L. (1999). A recurrent neural network that learns to count. *Connection Science*, *11*(1), 5–40.
- Rogers, T., & McClelland, J. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT Press.
- Rule, J. S., Tenenbaum, J. B., & Piantadosi, S. T. (2020). The child as hacker. *Trends in Cognitive Sciences*, *24*(11), 900–915.
- Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167.
- Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.
- Shi, L., & Griffiths, T. L. (2009). Neural implementation of hierarchical Bayesian inference by importance sampling. In Y. Bengio et al. (Eds.), *Advances in neural information*



- processing systems 22 (NIPS 2009)*. Curran Associates, Inc. [https://papers.nips.cc/paper\\_files/paper/2009/file/754dda4b1ba34c6fa89716b85d68532b-Paper.pdf](https://papers.nips.cc/paper_files/paper/2009/file/754dda4b1ba34c6fa89716b85d68532b-Paper.pdf)
- Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021, May 3–7). *Score-based generative modeling through stochastic differential equations* [Conference session]. 9th International Conference on Learning Representations, Virtual.
- Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24, 629–641.
- Wang, Z., Ku, A., Baldridge, J. M., Griffiths, T. L., & Kim, B. (2023). Gaussian process probes (GPP) for uncertainty-aware probing. In A. Oh et al. (Eds.), *Advances in neural information processing systems 36 (NIPS 2023)*. Curran Associates, Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/c8b100b376a7b338c84801b699935098-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/c8b100b376a7b338c84801b699935098-Paper-Conference.pdf)
- Yang, Y., & Piantadosi, S. T. (2022). One model for the learning of language. *Proceedings of the National Academy of Sciences, USA*, 119(5), Article e2021865119. <https://doi.org/10.1073/pnas.2021865119>