

Meta-learning as a bridge between neural networks and symbolic Bayesian models

R. Thomas McCoy and Thomas L. Griffiths

Open Peer Commentary on “Meta-learned models of cognition.” The version of record for this piece is available at <https://doi.org/10.1017/S0140525X24000116>.

Meta-learning is even more broadly relevant to the study of inductive biases than Binz et al. suggest: Its implications go beyond the extensions to rational analysis that they discuss. One noteworthy example is that meta-learning can act as a bridge between the vector representations of neural networks and the symbolic hypothesis spaces used in many Bayesian models.

Like many aspects of cognition, learning can be analyzed at multiple levels. At a high level (Marr’s “computational” level) we can model learning by providing an abstract characterization of the learner’s inductive biases: the preferences that the learner has for some types of generalizations over others (Mitchell, 1997). At a lower level, learning can be modeled by specifying the particular algorithms and representations that the learner uses to realize its inductive biases. For each of these levels, there are modeling traditions that have been successful: Rational analysis and Bayesian models are defined at the computational level, while neural networks are defined at the level of algorithm and representation. But how can we connect these different traditions? How can we work toward unified theories that bridge the divide between levels? In this piece, we agree with, and extend, Binz et al.’s point that meta-learning is a powerful tool for studying inductive biases in a way that spans levels of analysis.

Binz et al. describe how an agent can use meta-learning to derive inductive biases from its environment. This makes meta-learning well-suited for modeling situations where human inductive biases align with some problem that humans face—the situations that are well-covered by the paradigm of rational analysis (Anderson, 1990). As Binz et al. discuss, meta-learning can therefore be used to enable an algorithmically defined model (such as a neural network) to find the solution predicted by rational analysis, a procedure that bridges the divide between abstract rational solutions and specific algorithmic instantiations.

This direction laid out by Binz et al. is exciting. We argue that it can in fact be viewed as one special case within a broader space of possible lines of inquiry about inductive biases that meta-learning opens up. In the more general case, the Bayesian perspective allows us to define an inductive bias as a probability distribution over hypotheses. A neural network can meta-learn from data sampled from

this distribution, giving it the inductive bias in question. The distribution that is used could be drawn from (an approximation of) a human’s experience, in which case this framing matches the extension of rational analysis that Binz et al. advocate for. But it is also possible to use other approaches for defining this distribution, which can correspond to any probabilistic model. Since we can control probabilistic models, using a probabilistic model to define the distribution makes it possible to control the inductive biases that the meta-learned model ends up with (Lake, 2019; Lake and Baroni, 2023; McCoy et al., 2020). This allows us to take an inductive bias defined at Marr’s computational level and distill it into a neural network defined at the level of algorithm and representation.

Traditionally, certain types of inductive biases have been associated with certain types of algorithms and representations: The strong inductive biases of Bayesian models have generally been based on discrete, symbolic representations (e.g., Goodman et al., 2008), while neural networks use continuous vector representations Hinton et al. (1986) and have weak inductive biases. However, meta-learning enables us to separately manipulate inductive biases and representations, making it possible to model previously inaccessible combinations of representations and inductive biases. One noteworthy example is that we can use meta-learning to give symbolic inductive biases to a neural network, allowing us to study whether and how structured hypothesis spaces (of the sort often used in Bayesian models) can be realized in a system with continuous vector representations (the type of representation that is central in both biological and artificial neural networks). Thus, while Binz et al. note that meta-learning can be used as an alternative to Bayesian models, another use of meta-learning is in fact to expand the applicability of Bayesian approaches by reconciling them with connectionist models—thereby bringing together two successful research traditions that have often been framed as antagonistic (e.g., Griffiths et al., 2010; McClelland et al., 2010).

In our prior work, we have demonstrated the efficacy of this approach in the domain of language (McCoy and Griffiths, 2025). We started with a Bayesian model created by Yang and Piantadosi (2022), whose inductive bias is defined using a symbolic grammar. We then used meta-learning (specifically, MAML: Finn et al., 2017; Grant et al., 2018) to distill this Bayesian model’s prior into a neural network. The resulting system had strong inductive biases of the sort traditionally found only in symbolic models, enabling this system to learn formal linguistic patterns from small numbers of examples despite being a neural network, a class of systems that normally requires far more examples to learn such patterns. Additionally, the flexible neural implementation of this system made it possible to train it on naturalistic textual data, something that is intractable with the Bayesian model that we built on. Thus, meta-learning enabled the creation of a model that combined the complementary strengths of Bayesian and connectionist models of language learning.

These results show that inductive biases traditionally defined using symbolic Bayesian models can instead be realized inside a neural network. Therefore, symbolic inductive biases do not necessarily require inherently symbolic representations or algorithms. This demonstration provides one already-realized example of how meta-learning can advance our understanding of foundational questions about

how different levels of cognition relate to each other, in ways that go beyond the realm of rational analysis.

Financial support

This material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship under Grant No. 2204152 and the Office of Naval Research under Grant No. N00014-18-1-2873.

References

- Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.
- Binz, M., Dasgupta, I., Jagadish, A. K., Botvinick, M., Wang, J. X., and Schulz, E. (2024). Meta-learned models of cognition. *Behavioral and Brain Sciences*, 47:e147.
- Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning*.
- Goodman, N. D., Tenenbaum, J. B., Feldman, J., and Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32 1:108–54.
- Grant, E., Finn, C., Levine, S., Darrell, T., and Griffiths, T. (2018). Recasting gradient-based meta-learning as hierarchical Bayes. In *International Conference on Learning Representations*.
- Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364.
- Hinton, G. E., McClelland, J. L., and Rumelhart, D. E. (1986). Distributed representations. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, pages 77–109.
- Lake, B. M. (2019). Compositional generalization through meta sequence-to-sequence learning. *Advances in Neural Information Processing Systems*, 32.
- Lake, B. M. and Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman.

- McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., and Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14:348–56.
- McCoy, R. T., Grant, E., Smolensky, P., Griffiths, T. L., and Linzen, T. (2020). Universal linguistic inductive biases via meta-learning. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 737–743.
- McCoy, R. T. and Griffiths, T. L. (2025). Modeling rapid language learning by distilling Bayesian priors into artificial neural networks. *Nature Communications*, 16(1):4676.
- Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.
- Yang, Y. and Piantadosi, S. T. (2022). One model for the learning of language. *Proceedings of the National Academy of Sciences*, 119(5):e2021865119.