On the hazards of relating representations and inductive biases

Thomas L. Griffiths, Sreejan Kumar, and R. Thomas McCoy

Open Peer Commentary on "The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences"

The success of models of human behavior based on Bayesian inference over logical formulas or programs is taken as evidence that people employ a "language-of-thought" that has similarly discrete and compositional structure. We argue that this conclusion problematically crosses levels of analysis, identifying representations at the algorithmic level based on inductive biases at the computational level.

Over the last few decades probabilistic models of cognition, which explain human behavior in terms of Bayesian inference over a set of hypotheses, have been applied to a wide range of phenomena (Griffiths et al., 2010). But what does the success of a particular probabilistic model in capturing human behavior imply? Probabilistic models of cognition are typically defined at Marr's (1982) "computational" level, characterizing the abstract problems human minds have to solve and their ideal solutions. More precisely, they characterize the ideal solutions to inductive problems, where an agent has to draw conclusions that go beyond the available data. The content of a probabilistic model of cognition, expressed via the set of hypotheses and their prior probabilities, is a claim about the inductive biases that guide such inferences—those factors other than the data that influence the hypothesis the agent selects (Mitchell, 1997). Here, we argue that drawing conclusions that go beyond these inductive biases—and in particular, inferring support for specific cognitive processes and representations—can be problematic.

Inductive biases are at a different level of analysis from cognitive processes and representations, which Marr (1982) located at the "representation and algorithm" level. Representations and algorithms are notoriously underdetermined by observable data (Anderson, 1978). This underdetermination motivated Anderson (1990) to develop rational analysis, the approach adopted in almost all applications of probabilistic models of cognition. This approach explicitly focuses on abstract problems and their ideal solutions rather than the processes and representations that implement them (inspiring critiques, e.g., Jones and Love 2011). Probabilistic models need some way of representing hypotheses, but such representations do not necessarily guide human behavior. Rather, they are theoretical constructs that help scientists describe inductive biases.

Finding that a particular inductive bias seems to characterize human behavior places constraints on the representations and algorithms that might be involved, but those constraints rarely pick out a unique

solution. To give a simple example, consider the problem of learning a linear relationship between two variables. A probabilistic model identifies a set of hypotheses (e.g., all linear functions), defines a prior distribution over those hypotheses, and then performs Bayesian inference. This kind of solution could be implemented by an agent that explicitly represents a set of linear functions and uses an algorithm to update its beliefs about the posterior probability of each hypothesis as new data are observed (see, e.g., Sanborn et al. 2010). The behavior of this agent will match that of the ideal Bayesian model. However, an agent that seems quite different—a neural network with one hidden layer and a linear output function that updates its weights by a few iterations of gradient descent—will also produce an answer that matches that Bayesian model (assuming a Gaussian prior; see Santos 1996). Two very different representations and algorithms are consistent with the same computational-level account (for a real modeling example, see Feldman et al. 2009).

Quilty-Dunn et al. (2023) argue from the success of probabilistic models of cognition based on Bayesian inference over logical formulas and programs to the conclusion that people employ a similarly discrete and compositional "language-of-thought." This argument crosses levels of analysis in the same problematic way. What we are licensed to conclude from the success of these models is that logical formulas and programs are useful in characterizing human inductive biases for certain problems, not that humans use these representations when solving those problems. Any stronger conclusion seems particularly problematic in light of the recent successes of deep neural networks that Quilty-Dunn et al. mention, because these systems may not require discrete or compositional representations. Metalearning—training a system to perform a set of related tasks—provides a way to create neural networks with specific inductive biases, and has formal connections to learning a prior for Bayesian inference (Grant et al., 2018). Metalearning has been used to train neural networks to perform tasks characterized at the computational level by Bayesian models based on symbolic representations, such as theory-of-mind (Rabinowitz et al., 2018) and causal learning (Dasgupta et al., 2019). Analysis of the internal representations of related systems shows that they contain information that can be used to reconstruct appropriate posterior distributions (Mikulik et al., 2020). It thus seems plausible that such systems might produce behavior that is just as consistent with Bayesian inference over logical formulas and programs as that of humans.

The possible existence of deep neural networks that can be analyzed at the computational level in terms of Bayesian inference blocks strong conclusions about the language-of-thought, as the representations learned by these networks could emulate the associated behavior without requiring discreteness. Our investigations of networks trained by metalearning show that they can emulate human performance on abstract tasks without explicit representations of the relevant abstractions (Kumar et al., 2022). In some cases, deep neural networks succeed on abstract tasks by learning compositionally structured representations (McCoy et al., 2019), but these representations remain continuous, making them importantly different from the inherently discrete ones postulated in the language-of-thought hypothesis (Smolensky, 1988). Such results align with theoretical work showing that compositional behavior does not require discrete representations (Smolensky et al., 2022). Indeed, the best current models of language itself—which

is the prototypical example of a compositional domain (Pinker and Prince, 1988), as suggested by its use in the name *language-of-thought*—are deep networks that have continuous internal representations (e.g., Chowdhery et al. 2023).

Algorithms and representations may not be identifiable, but we can at least narrow down the equivalence class of possibilities through careful experimentation—behavioral work focused on response times and errors, neuroscientific studies of what the brain might be encoding, and computational simulations—designed to provide strong tests of alternative hypotheses. Until we can definitively do so, the fact that a discrete, compositional language-of-thought is useful as an abstract way of characterizing human inductive biases still allows the possibility that the actual representations and algorithms underlying human cognition may have a very different character.

Acknowledgment

We thank Tania Lombrozo for helpful comments.

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. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Office of Naval Research.

References

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, 85(4):249.

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N.

(2023). PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., Hughes, E., Battaglia, P., Botvinick, M., and Kurth-Nelson, Z. (2019). Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*.

Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4):752.

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.

Jones, M. and Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34(4):169–188.

Kumar, S., Correa, C. G., Dasgupta, I., Marjieh, R., Hu, M. Y., Hawkins, R., Cohen, J. D., Narasimhan, K., Griffiths, T., et al. (2022). Using natural language and program abstractions to instill human inductive biases in machines. *Advances in Neural Information Processing Systems*, 35:167–180.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W.H. Freeman.

McCoy, R. T., Linzen, T., Dunbar, E., and Smolensky, P. (2019). RNNs implicitly implement tensor-product representations. In *International Conference on Learning Representations*.

Mikulik, V., Delétang, G., McGrath, T., Genewein, T., Martic, M., Legg, S., and Ortega, P. (2020). Meta-trained agents implement Bayes-optimal agents. *Advances in Neural Information Processing Systems*, 33:18691–18703.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Pinker, S. and Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193.

Quilty-Dunn, J., Porot, N., and Mandelbaum, E. (2023). The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, 46:e261.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., and Botvinick, M. (2018). Machine theory of mind. In *International Conference on Machine Learning*, pages 4218–4227. PMLR.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117(4):1144.

Santos, R. J. (1996). Equivalence of regularization and truncated iteration for general ill-posed problems. *Linear Algebra and its Applications*, 236:25–33.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11(1):1–23.

Smolensky, P., McCoy, R., Fernandez, R., Goldrick, M., and Gao, J. (2022). Neurocompositional computing: From the central paradox of cognition to a new generation of AI systems. *AI Magazine*, 43(3):308–322.