Behavioral and Brain Sciences

# Is human compositionality meta-learned?

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | BBS-D-24-00113 |
| **Full Title:** | Is human compositionality meta-learned? |
| **Short Title:** | Meta-learned compositionality |
| **Article Type:** | Open Peer Commentary |
| **Corresponding Author:** | Michael Frank<br>Brown University<br>UNITED STATES |
| **Corresponding Author Secondary Information:** | |
| **Corresponding Author's Institution:** | Brown University |
| **Corresponding Author's Secondary Institution:** | |
| **First Author:** | Jacob Russin, PhD |
| **First Author Secondary Information:** | |
| **Order of Authors:** | Jacob Russin, PhD |
| | Sam Whitman McGrath, PhD |
| | Ellie Pavlick, PhD |
| | Michael J. Frank, PhD |
| **Order of Authors Secondary Information:** | |
| **Abstract:** | Recent studies suggest that meta-learning may provide an original solution to an enduring puzzle about whether neural networks can explain compositionality—in particular, by raising the prospect that compositionality can be understood as an emergent property of an inner-loop learning algorithm. We elaborate on this hypothesis and consider its empirical predictions regarding the neural mechanisms and development of human compositionality. |

**Authors of target article**: Marcel Binz, Ishita Dasgupta, Akshay Jagadish, Matthew Botvinick, Jane Wang, Eric Schulz

**Word counts**:

- Abstract: 60
- Main text: 996
- References: 944
- Entire text: 2123

**Commentary title**: Is human compositionality meta-learned?

**Commentary authors**: Jacob Russin, Sam Whitman McGrath, Ellie Pavlick, Michael J. Frank

**Institution:** Brown University

**Institutional mailing address:**

Metcalf Research Building
190 Thayer St
Providence, RI 02912

**Institutional phone number:** (401) 863-2727

**Email addresses:**

- Jacob Russin: jake_russin@brown.edu
- Sam Whitman McGrath: sam_mcgrath1@brown.edu
- Ellie Pavlick: ellie_pavlick@brown.edu
- Michael J. Frank: michael_frank@brown.edu

**Home page URLs:**

- Jacob Russin: https://jlrussin.github.io/
- Sam Whitman McGrath: https://scholar.google.com/citations?user=B3b7kAYAAAAJ&hl=en
- Ellie Pavlick: https://cs.brown.edu/people/epavlick/
- Michael J. Frank: http://ski.clps.brown.edu/

**60-word abstract**: Recent studies suggest that meta-learning may provide an original solution to an enduring puzzle about whether neural networks can explain compositionality—in particular, by raising the prospect that compositionality can be understood as an emergent property of an inner-loop learning algorithm. We elaborate on this hypothesis and consider its empirical predictions regarding the neural mechanisms and development of human compositionality.

Binz et al. (2023) review recent meta-learned models that can reproduce human-like compositional generalization behaviors (Lake & Baroni, 2023), but they stop short of endorsing meta-learning as a theoretical framework for understanding human compositionality. Here, we elaborate on this proposal, articulating the hypothesis that human compositionality can be understood as an emergent property of an inner-loop, in-context learning algorithm that is itself meta-learned.

Compositionality has played a key theoretical role in cognitive science since its inception (Chomsky, 1957), providing an explanation for human systematic and productive generalization behaviors. These phenomena are readily explained by the compositionality of classical cognitive architectures, as the design of their symbolic representations and structure-sensitive operations intrinsically guarantees that they can redeploy familiar constituents in novel constructions (Fodor & Pylyshyn, 1988). It has been argued that neural networks are in principle incapable of playing the same explanatory role because they lack these architectural features (Fodor & Pylyshyn, 1988; Marcus, 1998).

Much work has explored inductive biases that might encourage compositionality to emerge in neural networks (Russin, Jo, et al., 2020; Smolensky, 1990; Webb et al., 2024), but meta-learning offers an original solution to the puzzle. As Binz et al. (2023) emphasize, when an inner-loop, in-context learning algorithm emerges within the activation dynamics of a meta-learning neural network, it can have fundamentally different properties than the outer-loop algorithm. Thus, even if the outer-loop algorithm lacks these inductive biases, the network may nevertheless implement an emergent in-context learning algorithm that embodies them implicitly.

Lake and Baroni (2023) have shown that such an inner-loop algorithm can pass tests of compositionality that standard neural networks fail (Lake & Baroni, 2018). The question, then, is whether such networks can serve as explanatory models of human compositional generalization. Can we think of human compositionality as an emergent property of an inner-loop, in-context learning algorithm? How might we evaluate such a hypothesis? Here, we consider two independent aspects of this proposal: first, its implications for neural mechanisms, and second, for development.

One straightforward mechanistic prediction is that employing inner-loop, in-context learning mechanisms, rather than outer-loop learning mechanisms, should facilitate compositional generalization behaviors. Cognitive and computational neuroscience provide empirical support for this prediction. Cognitive control—the ability to overcome existing prepotent responses and to flexibly adapt to arbitrary goals (Miller & Cohen, 2001)—is an important capacity for human in-context learning. The neural mechanisms known to be involved in cognitive control, such as working memory, gating, and top-down modulation in the prefrontal cortex (Miller & Cohen, 2001; O'Reilly & Frank, 2006; Russin, O'Reilly, et al., 2020), are also thought to be essential to compositional abilities like inferring and applying rules (Calderon et al., 2022; Collins & Frank, 2013; Frank & Badre, 2012; Kriete et al., 2013), deductive and inductive reasoning (Crescentini et al., 2011; Goel, 2007), and processing complex syntax (Thompson-Schill, 2005). Thus, a shared set of neural mechanisms may underlie both in-context learning and compositionality in humans, lending support to the meta-learning hypothesis.

A second, independent prediction is a developmental one—that human compositional generalization abilities are themselves meta-learned over the course of development. Adults come into any psychological experiment equipped with a wealth of prior experience. The meta-learning hypothesis predicts that this includes experiences encouraging the adoption of more

compositional learning strategies (i.e., ones sensitive to implicit compositional structure). In general, children exhibit a developmental trajectory consistent with this hypothesis. Older children learn new tasks more efficiently (Bergelson, 2020), especially when these tasks involve cognitive capacities essential to in-context learning, such as working memory and executive functions (Munakata et al., 2012). Furthermore, children improve throughout development on tasks involving the composition of rules (Piantadosi & Aslin, 2016; Piantadosi et al., 2018).

Innate mechanisms or inductive biases may still be required to successfully meta-learn a compositional inner-loop algorithm in the first place. Indeed, studies in machine learning have shown that architecture seems to be an important factor in determining whether in-context learning capabilities emerge (Chan et al., 2022). Similarly, findings from cognitive and computational neuroscience have emphasized the importance of architectural features such as prefrontal gating mechanisms for the emergence of abstract representations that could mediate subsequent in-context generalization abilities (Collins & Frank, 2013; Frank & Badre, 2012; Kriete et al., 2013; Rougier et al., 2005). These inductive biases can also explain incidental hierarchical rule learning and generalization in infants (Werchan et al., 2015, 2016). Thus, a combination of innate architectural features and meta-learning experiences may be necessary for human compositionality to emerge.

The meta-learning datasets used in previous modeling efforts have typically been developmentally unrealistic because they have been contrived to engender narrow compositional generalization abilities that are specific to a particular type of task. Could meta-learning in less explicitly structured learning scenarios lead to the acquisition of broader compositional generalization abilities? This question deserves careful empirical study, but we may draw a preliminary insight from the success of large language models (Brown et al., 2020), which develop in-context learning abilities (von Oswald et al., 2023; Xie et al., 2022) that in some cases exhibit human-like compositionality (Webb et al., 2022; Wei et al., 2023; Zhou et al., 2022). Unlike models explicitly designed for meta-learning, large language models are trained to predict the next token on very large datasets of unstructured text. These datasets contain more language data than humans are exposed to in an entire lifetime (Linzen & Baroni, 2021), so future work will need to investigate what kinds of inductive biases are needed to improve their sample efficiency. However, these models provide proof of concept that neural networks can develop compositional in-context learning algorithms by training on relatively unstructured data.

Binz et al. (2023) shy away from a robust commitment to meta-learning as a theoretical framework, instead emphasizing its utility as a methodological tool. Here we have demonstrated how the meta-learning perspective on human compositionality can generate testable empirical hypotheses about underlying mechanisms and developmental trajectory. If such a research program bears fruit, it will elevate meta-learning from a useful tool to a novel cognitive theory.

**Competing interests:** None.

**Funding statement:** MJF is supported by ONR grant N00014-23-1-2792. EP and JR are supported by COBRE grant #5P20GM103645-10.

**References**


Bergelson, E. (2020). The Comprehension Boost in Early Word Learning: Older Infants Are Better Learners. *Child Development Perspectives*, *14*(3), 142–149. https://doi.org/10.1111/cdep.12373

Binz, M., Dasgupta, I., Jagadish, A., Botvinick, M., Wang, J. X., & Schulz, E. (2023). *Meta-Learned Models of Cognition*. https://doi.org/10.48550/ARXIV.2304.06729

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165; Issue arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Calderon, C. B., Verguts, T., & Frank, M. J. (2022). Thunderstruck: The ACDC model of flexible sequences and rhythms in recurrent neural circuits. *PLOS Computational Biology*, *18*(2), e1009854. https://doi.org/10.1371/journal.pcbi.1009854

Chan, S. C. Y., Santoro, A., Lampinen, A. K., Wang, J. X., Singh, A., Richemond, P. H., McClelland, J., & Hill, F. (2022). *Data Distributional Properties Drive Emergent In-Context Learning in Transformers* (arXiv:2205.05055; Issue arXiv:2205.05055). arXiv. http://arxiv.org/abs/2205.05055

Chomsky, N. (Ed.). (1957). *Syntactic structures*. Mouton & Co.

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering and generalizing task-set structure. *Psychological Review*, *120*(1), 190–229. https://doi.org/10.1037/a0030852

Crescentini, C., Seyed-Allaei, S., De Pisapia, N., Jovicich, J., Amati, D., & Shallice, T. (2011). Mechanisms of Rule Acquisition and Rule Following in Inductive Reasoning. *Journal of Neuroscience*, *31*(21), 7763–7774. https://doi.org/10.1523/JNEUROSCI.4579-10.2011

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Frank, M. J., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cerebral Cortex (New York, N.Y.: 1991)*, *22*(3), 509–526. https://doi.org/10.1093/cercor/bhr114

Goel, V. (2007). Anatomy of deductive reasoning. *Trends in Cognitive Sciences*, *11*(10), 435–441. https://doi.org/10.1016/j.tics.2007.09.003

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*, *110*(41), 16390–16395. https://doi.org/10.1073/pnas.1303547110

Lake, B. M., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In J. G. Dy & A. Krause (Eds.), *Proc. Of the 35th Intern. Conf. On Mach. Lear.* (Vol. 80, pp. 2879–2888). PMLR. http://proceedings.mlr.press/v80/lake18a.html

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, 1–7. https://doi.org/10.1038/s41586-023-06668-3

Linzen, T., & Baroni, M. (2021). Syntactic Structure from Deep Learning. *Annual Review of Linguistics*, *7*(1), null. https://doi.org/10.1146/annurev-linguistics-032020-051035

Marcus, G. F. (1998). Rethinking Eliminative Connectionism. *Cognitive Psychology*, *37*(3), 243–282. https://doi.org/10.1006/cogp.1998.0694

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Munakata, Y., Snyder, H. R., & Chatham, C. H. (2012). Developing Cognitive Control: Three Key Transitions. *Current Directions in Psychological Science*, *21*(2), 71–77. https://doi.org/10.1177/0963721412436807

O'Reilly, R. C., & Frank, M. J. (2006). Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. *Neural Computation*, *18*(2), 283–328. https://doi.org/10.1162/089976606775093909

Piantadosi, S., & Aslin, R. (2016). Compositional Reasoning in Early Childhood. *PLOS ONE*, *11*(9), e0147734. https://doi.org/10.1371/journal.pone.0147734

Piantadosi, S. T., Palmeri, H., & Aslin, R. (2018). Limits on composition of conceptual operations in 9-month-olds. *Infancy : The Official Journal of the International Society on Infant Studies*, *23*(3), 310–324. https://doi.org/10.1111/infa.12225

Rougier, N. P., Noelle, D., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal Cortex and the Flexibility of Cognitive Control: Rules Without Symbols. *Proceedings of the National Academy of Sciences*, *102*(20), 7338–7343.

Russin, J., Jo, J., O'Reilly, R. C., & Bengio, Y. (2020). Systematicity in a Recurrent Neural Network by Factorizing Syntax and Semantics. *Proceedings for the 42nd Annual Meeting of the Cognitive Science Society*, 7. https://cognitivesciencesociety.org/cogsci20/papers/0027/0027.pdf

Russin, J., O'Reilly, R. C., & Bengio, Y. (2020). Deep learning needs a prefrontal cortex. *Bridging AI and Cognitive Science (BAICS) Workshop, ICLR 2020*, 11.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1–2), 159–216. https://doi.org/10.1016/0004-3702(90)90007-M

Thompson-Schill, S. L. (2005). Dissecting the language organ: A new look at the role of Broca ' s area in language processing. In *Twenty-First Century Psycholinguistics* (1st ed., Vol. 1, pp. 1–18). Routledge.

von Oswald, J., Niklasson, E., Schlegel, M., Kobayashi, S., Zucchet, N., Scherrer, N., Miller, N., Sandler, M., Arcas, B. A. y, Vladymyrov, M., Pascanu, R., & Sacramento, J. (2023). *Uncovering mesa-optimization algorithms in Transformers* (arXiv:2309.05858). arXiv. https://doi.org/10.48550/arXiv.2309.05858

Webb, T., Frankland, S. M., Altabaa, A., Krishnamurthy, K., Campbell, D., Russin, J., O'Reilly, R., Lafferty, J., & Cohen, J. D. (2024). *The Relational Bottleneck as an Inductive Bias for Efficient Abstraction* (arXiv:2309.06629). arXiv. http://arxiv.org/abs/2309.06629

Webb, T., Holyoak, K. J., & Lu, H. (2022). *Emergent Analogical Reasoning in Large Language Models* (arXiv:2212.09196; Version 1). arXiv. http://arxiv.org/abs/2212.09196

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (arXiv:2201.11903). arXiv. https://doi.org/10.48550/arXiv.2201.11903

Werchan, D. M., Collins, A. G. E., Frank, M. J., & Amso, D. (2015). 8-Month-Old Infants Spontaneously Learn and Generalize Hierarchical Rules. *Psychological Science*, *26*(6), 805–815. https://doi.org/10.1177/0956797615571442

Werchan, D. M., Collins, A. G. E., Frank, M. J., & Amso, D. (2016). Role of Prefrontal Cortex in Learning and Generalizing Hierarchical Rules in 8-Month-Old Infants. *The Journal of Neuroscience*, *36*(40), 10314–10322. https://doi.org/10.1523/JNEUROSCI.1351-16.2016

Xie, S. M., Raghunathan, A., Liang, P., & Ma, T. (2022). *An Explanation of In-context Learning as Implicit Bayesian Inference* (arXiv:2111.02080). arXiv. http://arxiv.org/abs/2111.02080

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., & Chi, E. (2022). *Least-to-Most Prompting Enables Complex Reasoning in Large Language Models* (arXiv:2205.10625). arXiv. http://arxiv.org/abs/2205.10625