Constructing Deep Concepts through Shallow Search

Bonan Zhao[1], **Christopher G. Lucas**[2], **Neil R. Bramley**[3]

[1]Department of Computer Science, Princeton University
[2]School of Informatics, University of Edinburgh
[3]Department of Psychology, University of Edinburgh
bnz@princeton.edu, c.lucas@ed.ac.uk, neil.bramley@ed.ac.uk

Introduction

In 1675, Isaac Newton wrote in a letter to Robert Hooke: "If I have seen further, it is by standing on the shoulders of giants." This remark elegantly captured the incremental nature of human learning. We seem to extend and repurpose existing knowledge to create new and more powerful ideas. Here, we refer to this ability as "bootstrap learning", and particularly focus on bootstrap learning inductive concepts. It is helpful to think of bootstrap learning as a solution to how cognitively-bounded reasoners (Anderson 1990; Griffiths, Lieder, and Goodman 2015) grasp complex environmental dynamics that are far beyond their initial capacity (Figure 1a). Human cognition is constrained by limited time, memory, communication means, etc, but the underlying learning and generalization problems posed by the environment can be unboundedly complex. Rather than the typical machine learning approach to scaling, overwhelming every problem with a larger architecture, more data, and more training cycles, bootstrap learning offers a way to maximize the reach and potential of a learner with a fixed search and representational budget, by searching 'locally' and recursively to extend their existing knowledge (Bramley et al. 2023).

Successful bootstrap learning is reliant on discovering the right sub-concepts that "carve nature at its joints" (Plato 1952/370BC). This depends crucially on how and in what order evidence is processed. We report an experiment that demonstrates people construct drastically different causal concepts and generalizations upon seeing the same set of evidence presented in different orders (Zhao, Lucas, and Bramley 2023). As illustrated in Figure 1b, participants observed six examples of animated evidence generated by a ground-truth rule $R' \leftarrow \text{stripe}(A) \times R - \text{spot}(A)$. In the *construct* condition, people first saw evidence consistent with $R' \leftarrow \text{stripe}(A) \times R$ (trials 1-3), and then saw an additional batch of evidence introducing the spots (trials 4-6). In the *de-construct* condition, people first saw trials 4-6, and then an trials 1-3. The two groups thus had access to identical information, just in different batch orders. Most participants in the *construct* condition could infer the multi-captive sub-concept and subsequently discovered the ground truth. People in the *de-construct* condition, however, struggled to come up with a simple functional relationship that best explains what they saw; they came up with complex disjunctive rules, or simply reported "I don't know!". This demonstrates that (1) human learning benefits from facilitatory curricula, where having the chance to create a valid sub-concept is crucial for constructing more complex compound concepts, and (2) human inductive inference is susceptible to learning traps where, once formed, inappropriate sub-concepts are hard to shake and can impair subsequent inductive learning.

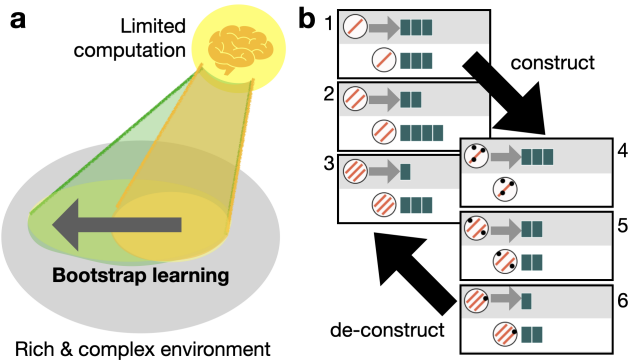

Figure 1: a. Illustration of bootstrap learning: Learners with limited computation construct more complex ideas by reusing previous findings. b. Stimuli in Zhao, Lucas, and Bramley (2023).

Resource-rational Library Learning

How can we design artificial systems that build representations, and make generalizations, like those exhibited by people? We argue that there are two key components: (1) a structured representational substrate, and (2) an effective cache-and-reuse mechanism.

Rich, structured representations, like symbolic programs, encode human-like regularities in inference, and can achieve human-like performance given human-level input data (Lake, Salakhutdinov, and Tenenbaum 2015). In addition to striking a balance between structured symbolic knowledge and probabilistic inference, thanks to methods like probabilistic context-free grammars, the space of possible pro-

grams generated by existing symbolic knowledge can be open-ended, like how human concept spaces are (e.g., Goodman et al. 2008). In particular, we draw inspirations from Bayesian library learning, a class of methods aiming to learn shareable and reusable sub-programs, or 'libraries', that facilitate fast and flexible learning (Dechter et al. 2013; Ellis et al. 2021; Liang, Jordan, and Klein 2010). These methods relax the context-free assumption used in traditional Bayesian symbolic models, and jointly infers both the posterior over concept 'programs', and a latent library that defines this posterior. This notion of concept libraries is attracting increasing attention across cognitive science and generative AI (Bowers et al. 2023; Tian et al. 2020; Wang et al. 2023; Wong et al. 2022).

We argue that human-like library learning are constrained by the amount of resources available to people (Anderson 1990). In our model, we introduced a dynamic concept library to a classic Bayesian symbolic learning framework, powered by an adaptor grammar representation, a generalization of probabilistic context-free grammars (Johnson et al. 2007). Different from standard library learning approaches, this allows us to assume a shallow search depth cap, mimicking cognitively-bounded learners. Different from naive Bayesian learners, this model can cache learned programs into its library, and later reuse these programs to construct more complex programs. Therefore, this model can construct deeply nested programs that go far beyond its search depth constraints, and will succeed in doing so under facilitatory learning curricula. Our model predicts not only when people succeed at learning complex concepts, but also when people fail to do so. Quantitative fits with human behavioral data also showed that this rational library learning model best matched participants generalizations.

Broader Implications

This resource-rational library learning framework offers a computational account for why human learning is usually incremental and path-dependent. Computational constraints of human cognition determine that we can only process limited information, draw limited number of samples, and search a limited space at a time. However, bootstrap learning mechanisms enable us to reach beyond our grasp, and explore an ever-richer space of possibilities via principled cache-and-reuse. The process of caching the current conceptual constructs and later reusing them to form more advanced ideas may also give rise to the hierarchical structure observed in human conceptual systems.

The possibilities afforded by bootstrapping are not a 'free lunch'. As illustrated by behavioral experiments, bounded agents fall easily to learning traps (Rich and Gureckis 2018) in inductive inference. The fact that people draw systematically different conclusions after seeing the same evidence is worth being taken seriously if we want to design more human-like learning systems. Synthesizing the kinds of subconcepts people create in the process of reaching a complex learning target may be more important than matching final learning performance, because our conceptual systems are built from these interacting auxiliary concepts and their dynamics. Identifying how and where people diverge is also key for designing personalized learning algorithms like automatic teaching assistants and consultation agents, especially in face of increasing interest in aligning human and artificial learning systems (Sucholutsky et al. 2023).

The fact that human learning benefits from simple to complex curricula has inspired the training of many artificial systems, from early work manipulating the system's capacity (Elman 1993), to carefully designed training curricula for deep neural networks (Bengio et al. 2009) and neuro-symbolic models (Mao et al. 2019). It is worth noting that those models still require substantial training, while human concept learning, as demonstrated in the experiments and modeled by Zhao, Lucas, and Bramley (2023), can be driven by a handful of observations. Structured representations are still strong candidates for data-efficient, human-like learning algorithms.

In sum, we propose bootstrap learning as a computational account for why human learning is modular and incremental, and identify key components of bootstrap learning that allow artificial systems to learn more like people. We offer both a computational modeling framework and behavioral evidence that highlights the double-edged sword of bootstrap learning, calling for the importance of taking resource constraints, diverse learning outcomes, and social aspects into account in designing increasingly human-like artificial systems.

Acknowledgments

This work was supported by an EPSRC New Investigator Grant (EP/T033967/1) to Bramley and Lucas.

References

Anderson, J. R. 1990. *The Adaptive Character of Thought*. Psychology Press.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Bowers, M.; Olausson, T. X.; Wong, L.; Grand, G.; Tenenbaum, J. B.; Ellis, K.; and Solar-Lezama, A. 2023. Top-down synthesis for library learning. *Proceedings of the ACM on Programming Languages*, 7(POPL): 1182–1213.

Bramley, N. R.; Zhao, B.; Quillien, T.; and Lucas, C. G. 2023. Local search and the evolution of world models. *Topics in Cognitive Science*.

Dechter, E.; Malmaud, J.; Adams, R. P.; and Tenenbaum, J. B. 2013. Bootstrap learning via modular concept discovery. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Ellis, K.; Wong, C.; Nye, M.; Sablé-Meyer, M.; Morales, L.; Hewitt, L.; Cary, L.; Solar-Lezama, A.; and Tenenbaum, J. B. 2021. DreamCoder: Bootstrapping inductive program synthesis with wake-sleep library learning. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, 835–850.

Elman, J. L. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1): 71–99.

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): 108–154.

Griffiths, T. L.; Lieder, F.; and 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.

Johnson, M.; Griffiths, T. L.; Goldwater, S.; et al. 2007. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. *Advances in neural information processing systems*, 19: 641.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338.

Liang, P.; Jordan, M. I.; and Klein, D. 2010. Learning programs: A hierarchical Bayesian approach. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 639–646.

Mao, J.; Gan, C.; Kohli, P.; Tenenbaum, J. B.; and Wu, J. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *International Conference on Learning Representations*.

Plato. 1952/370BC. *Phaedrus*, volume 1. Cambridge University Press.

Rich, A. S.; and Gureckis, T. M. 2018. The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology: General*, 147(11): 1553–1570.

Sucholutsky, I.; Muttenthaler, L.; Weller, A.; Peng, A.; Bobu, A.; Kim, B.; Love, B. C.; Grant, E.; Achterberg, J.; Tenenbaum, J. B.; et al. 2023. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*.

Tian, L.; Ellis, K.; Kryven, M.; and Tenenbaum, J. 2020. Learning abstract structure for drawing by efficient motor program induction. *Advances in Neural Information Processing Systems*, 33: 2686–2697.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Wong, C.; McCarthy, W. P.; Grand, G.; Friedman, Y.; Tenenbaum, J. B.; Andreas, J.; Hawkins, R. D.; and Fan, J. E. 2022. Identifying concept libraries from language about object structure. *arXiv preprint arXiv:2205.05666*.

Zhao, B.; Lucas, C. G.; and Bramley, N. R. 2023. A model of conceptual bootstrapping in human cognition. *Nature Human Behaviour*, 1–12.