From simple to complex: shared learning dynamics in humans and neural networks.

Jirko Rubruck[†] **(jirko.rubruck@stx.ox.ac.uk)**
Oxford Department of Experimental Psychology, Anna Watts Building, Woodstock Rd
Oxford, Oxfordshire OX2 6GG United Kingdom

Alice Zhang[†] **(alice.zhang@psy.ox.ac.uk)**
Oxford Department of Experimental Psychology, Anna Watts Building, Woodstock Rd
Oxford, Oxfordshire OX2 6GG United Kingdom

Satwik Bhattamishra (satwik.bmishra@stx.ox.ac.uk)
Oxford Department of Computer Science, Wolfson Building, Parks Rd
Oxford, Oxfordshire OX1 3QG United Kingdom

Christopher Summerfield (christopher.summerfield@psy.ox.ac.uk)
Oxford Department of Experimental Psychology, Anna Watts Building, Woodstock Rd
Oxford, Oxfordshire OX2 6GG United Kingdom

[†] These authors contributed equally to this work.

Abstract

Deep neural networks demonstrate a well-documented simplicity bias—the tendency to learn simple functions before acquiring more complex ones. This bias towards simplicity is thought to enable overparameterized models to successfully generalize to unseen data rather than overfitting to examples seen in training. Complementary work in psychology has demonstrated human simplicity biases in several domains. Here, we aim to unite these two streams by comparing human and neural network simplicity biases side-by-side in a Boolean classification task. We demonstrate that both humans and models initially learn simple rules before mastering a more complex function. We also provide evidence that human learners rely on the simple functions they learned early on to classify out-of-distribution examples, suggesting that dynamical simplicity biases are important for generalization.

Keywords: simplicity bias; learning dynamics; category learning; neural networks; generalization

Introduction

Deep neural networks start by learning simple, often linear models of the data (Kalimeris et al., 2019; Hu et al., 2020; Rubruck et al., 2024) and the complexity of the learned function increases with training (Bhattamishra et al., 2023; Saxe et al., 2019). A qualitatively similar phenomenon has been documented in human category learning, where cardinal rules are learned before exceptions (Nosofsky et al., 1994). The current work aims to explicitly connect the dynamical nature of simplicity biases in biological and artificial learning. To this end, we adapt a Boolean category learning task from Budiono et al. (2024). We theorize that humans and neural networks will be biased to learn simple functions early in training. Because simplicity biases are thought to be foundational for generalization in artificial learning (Bhattamishra et al., 2023; Kalimeris et al., 2019; Rahaman et al., 2019), we additionally hypothesize that human learners rely on simple functions they learned early on to generalize to out-of-distribution problems.

Methods

Behavioral Experiment

Boolean classification task. Participants ($N = 87$) were recruited via Prolific and learned to classify bee stimuli (Fig 1A). Each bee could be identified by a combination of three binary features: wing length, body pattern, and leg number. In a design inspired by Shepard et al. (1961), participants were tasked with learning classifications in which a simple rule based on a single feature (e.g. short wings-good, long wings-bad) could achieve above-chance accuracy, but could not correctly classify all stimuli. In learning trials, participants classified a bee and then received feedback on their choice (Fig 1B, left). Participants completed 48 blocks of learning in which the eight unique bees were shown in randomized order. After learning, participants completed a generalization phase with no feedback where each bee had one feature hidden (Fig 1B, right). Participants saw each generalization stimulus twice.

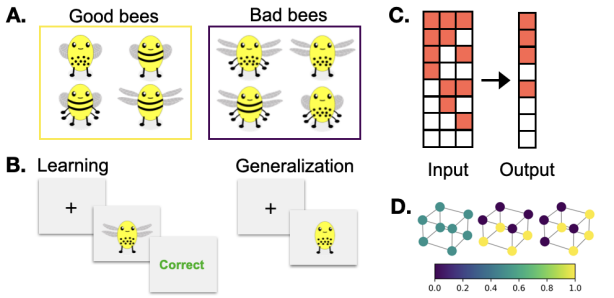

Figure 1: A. An example task learned by human participants. Notice that a simple rule based on one feature can be used to achieve above-chance accuracy. B. In learning trials, participants categorize bees with feedback. In generalization trials, each bee has one feature hidden and feedback is absent. C. Binary inputs and outputs are used to train networks on an abstracted version of the classification task. D. Cartoon sketch of simple to complex learning. Each point represents a 3-d stimuli and color represents its learned label. The left cube represents responding 'good bee' or 'bad bee' with equal probability. The middle cube represents reliance on a single feature, and the rightmost represents using the true labeling function from A.

Assessment of simplicity We chose a Boolean classification task in part due to the tools available to quantify the complexity of such functions (O'Donnell, 2021). We focus on sensitivity, a measure of how much the output of a Boolean function changes in response to "small" input perturbations. Intuitively, a simple function has low sensitivity because similar inputs are likely to have the same output. Formally, given inputs $\mathbf{x}_i \in \{0,1\}^n$ the sensitivity of a Boolean function for a particular input is defined as $s(f,\mathbf{x}) = \sum_{j=1}^{n} \mathbb{I}[f(\mathbf{x}) \neq f(\mathbf{x}^{\oplus j})]$. Here, $\mathbf{x}^{\oplus j}$ denotes the input that has a bit flip at position j and $\mathbb{I}$ the indicator function.

Using an HMM to infer decision functions. To assess the complexity of human decision functions, we need to infer these function from participants' choice data. In particular, we need to be able to distinguish between random behavior, which is expected at the start of learning, and the use of a very complex labeling function. To achieve this, we fit a Hidden Markov Model (HMM) to each participant's choices. This analysis assumes that each choice is generated by an underlying decision function, and that participants transition between distinct decision functions throughout learning. Therefore, each hidden state of the HMM represents a decision function, and its emission probabilities describe the probability of labeling each stimulus as good or bad. HMMs were fit for each participant using expectation-maximization, and used to infer the decision function underlying the participants choice on every trial. The number of hidden states (distinct decision functions)

was determined via model comparison using the Akaike information criterion to penalize model complexity.

Modeling

We trained a two-layer feed-forward neural network to perform our classification task. The data set consists of 3 dimensional Boolean vectors $\mathbf{x}_i \in \{0,1\}^3$. We used the same labeling functions learned by human participants for model training. Our networks contain a single hidden layer of dimensions 20 with ReLU nonlinearities, and an output unit with sigmoid activation. We train networks with squared error-loss and Adam optimizer with a learning rate of .02 and small initial weights.

Results

Learning functions of increasing complexity

We first assess whether participants learn functions of increasing complexity by looking at a heuristic measure, reliance on a simple rule based on one feature. If participants follow a simple to complex trajectory, they should learn stimuli that are correctly classified by the simple rule (central stimuli) faster than those that are not (peripheral stimuli; nomenclature adapted from Nosofsky et al. (1994)). Because some labeling functions have more than one valid simple rule (e.g. it is possible to classify above chance by relying on just wing number, or just body pattern), central stimuli are defined as those that are correctly classified by *all* valid simple rules. We find that participants learn central examples more quickly (Fig. 2A), which is in line with work by (Nosofsky et al., 1994). Surprisingly, we find that this same behavior is also displayed by simple neural networks trained from scratch.

Next, we look at another, more generalizable measure for function complexity over learning, Boolean sensitivity (see methods). To do so, we fit an HMM for each participant to infer the labeling function that generates their behavior on each trial. We then compute the sensitivity of that labeling function, and find that sensitivity increases over the course of learning (Fig. 2B). We replicate the same result in our neural network models by directly computing the sensitivity of model outputs.

Simplicity and generalization

We ask if participants rely on the simple rule they learned initially when generalizing to out-of-distribution stimuli. To do so, we focus on problems where participants could have initially learned a simple rule based on any of the three features. Additionally, we look only at participants whose performance exceeded that of the simple rule by the end of learning ($> 87.5\%$ accuracy in the last five blocks). For this analysis, we had $n = 20$ participants who learned this particular rule and reached the performance criterion. We find that reliance on a feature during learning predicts reliance on that same feature in generalization (Fig. 3). Learning reliance is average accuracy in learning on stimuli that can be correctly classified by a simple rule based on the relevant feature. Generalization reliance is the proportion of generalization trials—where the relevant feature is present—in which participants make choices that align with the same simple rule.

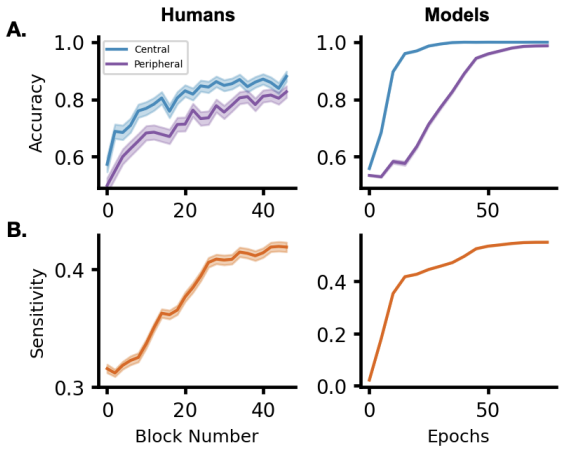

Figure 2: Humans and neural network models learn functions of increasing complexity. Values are averaged over every other block for humans and every five epochs for models. Model results are averaged over 1000 runs. A. Humans and models learn stimuli that are correctly classified by valid simple rules (central) faster than those that are not (peripheral). B. Sensitivity of the Boolean functions used by humans and models increases over the course of learning.

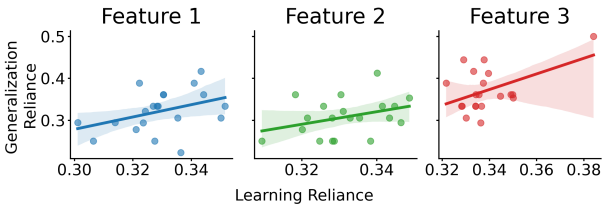

Figure 3: The extent to which participants rely on a feature during learning predicts their reliance on that feature in generalization. Generalization reliance and learning reliance were normalized to sum to 1 across features for each participant.

Discussion

In this work, we show that humans and neural network models learn functions of increasing complexity. We provide evidence that humans rely on the simple rule that is initially learned when generalizing, suggesting that dynamical simplicity biases may influence human generalization. Future work will assess whether our understanding of simplicity biases in learning can be leveraged to design more effective curricula.

Acknowledgments

This work was funded by a Wellcome Trust Discovery Award (227928/Z/23/Z) to C.S., a UKRI ESRC Grand Union Doctoral training partnership stipend awarded to J.R., and a Clarendon fund scholarship to A.Z.

References

Bhattamishra, S., Patel, A., Kanade, V., & Blunsom, P. (2023, July). *Simplicity Bias in Transformers and their Ability to Learn Sparse Boolean Functions.* arXiv. (arXiv:2211.12316)

Budiono, R., Hartley, C., & Gureckis, T. M. (2024). How does social learning affect stable false beliefs? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(0). Retrieved 2025-04-01, from https://escholarship.org/uc/item/3sw1d8m1

Hu, W., Xiao, L., Adlam, B., & Pennington, J. (2020). The Surprising Simplicity of the Early-Time Learning Dynamics of Neural Networks. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 17116–17128). Curran Associates, Inc.

Kalimeris, D., Kaplun, G., Nakkiran, P., Edelman, B., Yang, T., Barak, B., & Zhang, H. (2019). SGD on Neural Networks Learns Functions of Increasing Complexity. In *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79. (Place: US Publisher: American Psychological Association) doi: 10.1037/0033-295X.101.1.53

O'Donnell, R. (2021, May). *Analysis of Boolean Functions.* arXiv. Retrieved 2025-04-04, from http://arxiv.org/abs/2105.10386 (arXiv:2105.10386 [cs]) doi: 10.48550/arXiv.2105.10386

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., ... Courville, A. (2019, May). On the Spectral Bias of Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 5301–5310). PMLR. (ISSN: 2640-3498)

Rubruck, J., Bauer, J. P., Saxe, A., & Summerfield, C. (2024, June). *Early learning of the optimal constant solution in neural networks and humans.* arXiv. Retrieved 2025-04-04, from http://arxiv.org/abs/2406.17467 (arXiv:2406.17467 [cs]) doi: 10.48550/arXiv.2406.17467

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019, June). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1820226116

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42. (Place: US Publisher: American Psychological Association) doi: 10.1037/h0093825