Humans and Neural Networks Show Similar Patterns of Transfer and Interference in a Continual Learning Task

Eleanor Holton (eleanor.holton@psy.ox.ac.uk)
University of Oxford, Oxford, UK

Lukas Braun (lukas.braun@psy.ox.ac.uk)
University of Oxford, Oxford, UK

Jessica A.F. Thompson (jessica.thompson@psy.ox.ac.uk)
University of Oxford, Oxford, UK

Christopher Summerfield (christopher.summerfield@psy.ox.ac.uk)
University of Oxford, Oxford, UK
DeepMind, London, UK

Abstract:

When learning a new task, old task knowledge can incur both benefits (transfer) and costs (interference). How biological and artificial agents trade off these costs and benefits is an unsolved computational problem. One clue comes from recent analysis of learning in artificial neural networks (ANNs), which paradoxically show strongest interference for new tasks that are of intermediate levels of similarity to previous tasks (relative to those that are very similar or very different). Here, we directly compare this effect in humans and ANNs. In two successive tasks (A and B) humans and ANNs learned to map stimuli onto a continuously-valued circular output in two distinct contexts, where the outputs across contexts were related by a rule. By varying the similarity between task rules, we found that in both humans and ANNs, more similar tasks led to faster learning on task B, but more dissimilar tasks led to lower interference in memory of A. These results point to key parallels between humans and ANNs in continual learning settings, whereby task similarity promotes shared representations enabling faster acquisition of the new task, while task dissimilarity leads to lower catastrophic interference by invoking distinct representations.

Keywords: continual learning; catastrophic interference; transfer.

Introduction

Continual learning is the ability to learn new tasks while retaining good performance on previous tasks. It remains a significant challenge for artificial neural networks, whereas biological systems show an impressive ability to learn continuously throughout their lives (McCloskey & Cohen, 1989; Flesch et al. 2023). Here, we show a surprising case in which the patterns of transfer and interference shown by humans and neural networks seem to be qualitatively similar in nature.

Recent work in machine learning has revealed that the interference incurred by a new task depends non-monotonically on its similarity to past tasks, with very similar and very different tasks incurring the lowest costs (a phenomenon that has been dubbed "Maslow's Hammer"; see Lee 2021, 2022; Ramasesh 2021). This may be a natural consequence of the need to trade off transfer and interference: agents attempt to solve moderately similar tasks using existing representations (leading to faster learning but corrupting the representation for previous tasks) while forming new representations when tasks are sufficiently dissimilar. Do humans do the same? It has been proposed that continual learning in humans is enabled by forming separate neural task representations when major environmental changes are encountered (Gershman et al. 2015; Flesch et al. 2018; Pisupati & Niv, 2022).

Human Behaviour

To study Maslow's hammer in humans, we recruited 3 cohorts of participants (n=50 in each) to perform a task that involved learning to map discrete inputs ("plants") onto positions on a ring ("location") in two distinct contexts ("seasons"; see **Fig.1**). On every trial, a plant image was presented within a circular dial under a question indicating the particular season (**Fig.1A**). Participants dragged the dial handle to indicate the plant's location in that season, receiving supervised feedback on every trial. The relationship between a plant's locations in the two seasons was an offset of 60° in task A. In task B, the offset was shifted for new stimuli by 0° ("same rule" group), 60° ("near rule") or 180° ("far rule"; **Fig.1B**). After participants in each group were trained on both tasks sequentially, they were re-tested on the locations of the stimuli from task A, this time without feedback (**Fig.1C**).

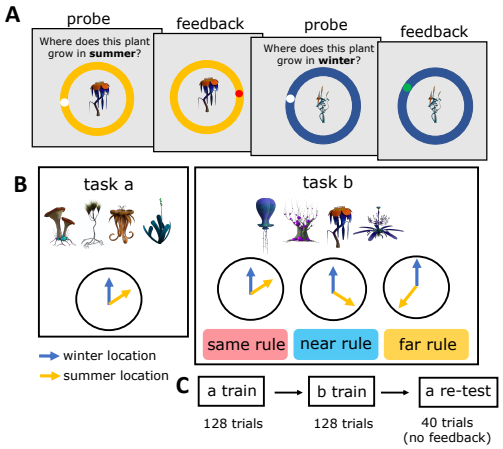

Figure 1: Study Design. **(A)** Two example trials. **(B)** Schematic of the sequential tasks **(C)** Study timeline

Results. Participants were faster at learning about new stimuli in task B when the seasonal rule remained consistent with task A (mean error on the first half of training; same vs. far: Mann-Whitney U=996.0.0, p=0.002; same vs. near: U=1101.0, p=0.002; far vs. near n.s.; see **Fig. 2B**), although all groups achieved the same performance on both tasks by the end of training. Memory interference was quantified as the difference in error between the second half of task A training, and the retest period on A stimuli (following task B training). Participants who learned a new task with the most dissimilar rule showed lower memory interference on the previous task than those for whom the rule remained similar across tasks (same-rule>far-rule: U=577.0, p=0.016; near-rule>far rule: U=551.0, p=0.040; near-rule>same-rule: n.s.; **Fig.2C**).

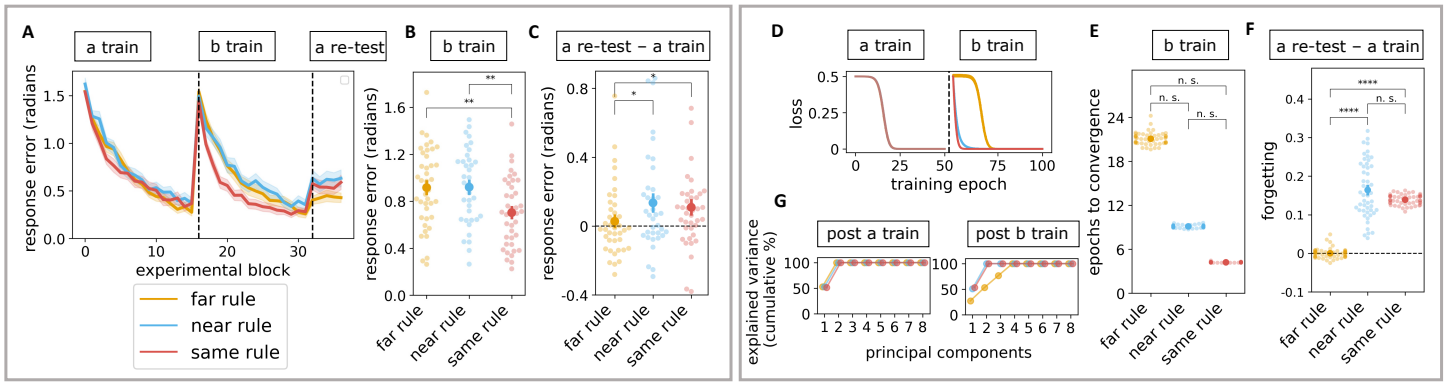

Figure 2: Human results (left) **(A)** Mean error over successive tasks **(B)** Task B error **(C)** Interference in task A performance; ANN results (right). **(D)** Learning loss for successive tasks **(E)** Speed of learning task B **(F)** Interference on task A **(G)** Principal components in network hidden layer activity across all inputs after training on A (left) and B (right)

ANN Behaviour

Setup and results. Networks were two-layer feed-forward linear networks with 20 hidden units to map one-hot vector inputs onto Cartesian coordinates for the corresponding angle. They were trained using online stochastic gradient descent (single trial batches) presented in participant-matched stimulus order. Network weights were not reset between the tasks.

Consistent with human participants, learning on task B was fastest when the rule linking the two features remained the same across tasks (median convergence time during task B: same-rule M=4.20, near-rule M=9.14, far-rule M=21.11, U<0.00 and p<0.001 for all three contrasts; **Fig.2E**). Also like humans, catastrophic forgetting was lower in the far-rule condition than in either the same- or near-rule conditions (same-rule M=0.1372, near-rule M=0.1536, far-rule M=0.0001; far vs. same: U=2500, p<0.001; far vs. near: U=2498.0, p<0.001; same vs near: n.s.; **Fig.2F**).

Notably, ANNs unlike human participants also displayed sharp learning costs when required to learn the far-rule condition. For neural networks, this is explained by critical differences in the hidden layer representations associated with the two tasks. Networks learning a rule which is an 180° reversal of the original rule suffer from being fixed in a saddle point during initial learning (see reversal learning case in Braun et al. 2022). This causes the network to eventually reach a task solution by using different representations for the two tasks. **Figure 2G** shows the number of principal components describing activity in the hidden layer when the ANN is probed on all stimuli across both tasks (at minimum two p.c. are needed to solve a single task). Networks in the same-rule and near-rule conditions use only two components across both tasks (recycling the same task representation), while networks in the far-rule condition converge with double the components, reflecting separate representations for the two tasks. This explains the longer learning times on task B but minimal interference on the original task.

Comparison with lazy regime Previous work has shown that ANNs initialized with large weights ("*lazy*" regime) converge on high-dimensional solutions that fail to capture the environmental statistics (Chizat et al. 2019, Flesch et al. 2022). While networks initialized with large weights reach equal performance on the task, they show no differences between conditions (**Fig.3**), suggesting these phenomena only apply when learners carry information about the structural rule.

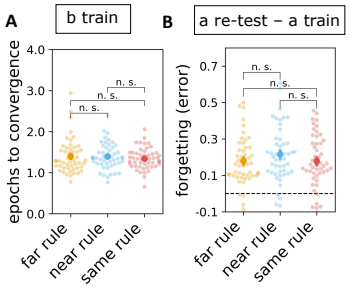

Figure 3: ANNs trained in the lazy regime show no differences in learning of task B (left) or interference on task A (right)

Discussion

We present evidence that both humans and ANNs learning successive tasks can benefit from task similarity when acquiring new knowledge ('transfer'), alongside reduced forgetting when tasks are dissimilar ('interference'). These findings suggest task similarity plays a key role in determining the trade-off between transfer and interference for both ANNs and humans in structured environments.

Acknowledgments

This work was supported by the Wellcome Trust (award 222347/Z/21/Z to E.H.).

References

Braun, L., Dominé, C., Fitzgerald, J., & Saxe, A. (2022). Exact learning dynamics of deep linear networks with prior knowledge. *Advances in Neural Information Processing Systems*, *35*, 6615–6629.

Chizat, L., Oyallon, E., & Bach, F. (2019). On Lazy Training in Differentiable Programming. *Advances in Neural Information Processing Systems*, *32*.

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, *115*(44), E10313–E10322.

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, *110*(7), 1258-1270.e11.

Flesch, T., Nagy, D. G., Saxe, A., & Summerfield, C. (2023). Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals. *PLOS Computational Biology*, *19*(1), e1010808.

Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, *5*, 43–50.

Lee, S., Goldt, S., & Saxe, A. (2021). Continual Learning in the Teacher-Student Setup: Impact of Task Similarity. *Proceedings of the 38th International Conference on Machine Learning*, 6109–6119.

Lee, S., Mannelli, S. S., Clopath, C., Goldt, S., & Saxe, A. (2022). Maslow's Hammer in Catastrophic Forgetting: Node Re-Use vs. Node Activation. *Proceedings of the 39th International Conference on Machine Learning*, 12455–12477.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 24, pp. 109–165). Academic Press.

Pisupati, S., & Niv, Y. (2022). The challenges of lifelong learning in biological and artificial systems. *Trends in Cognitive Sciences*, *26*(12), 1051–1053.

Ramasesh, V. V., Dyer, E., & Raghu, M. (2020). *Anatomy of Catastrophic Forgetting: Hidden Representations and Task Semantics* (arXiv:2007.07400).