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

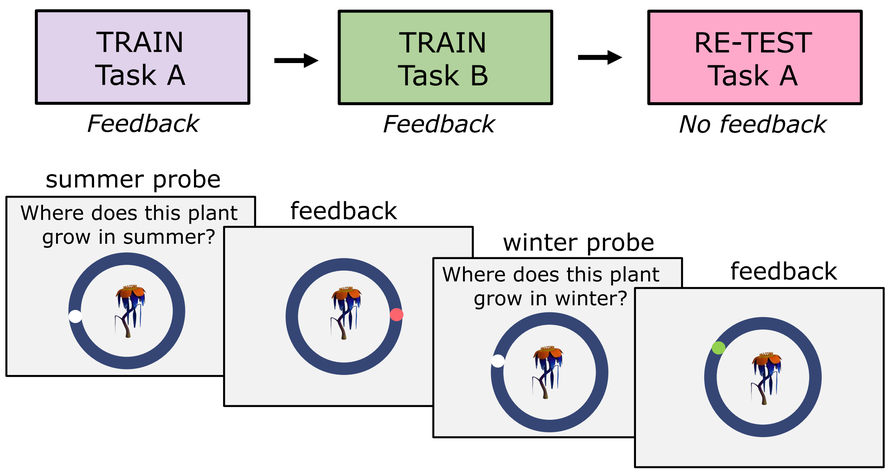
Eleanor Holton[1], Lukas Braun[1], Jessica A.F. Thompson[1], Christopher Summerfield[1,2]

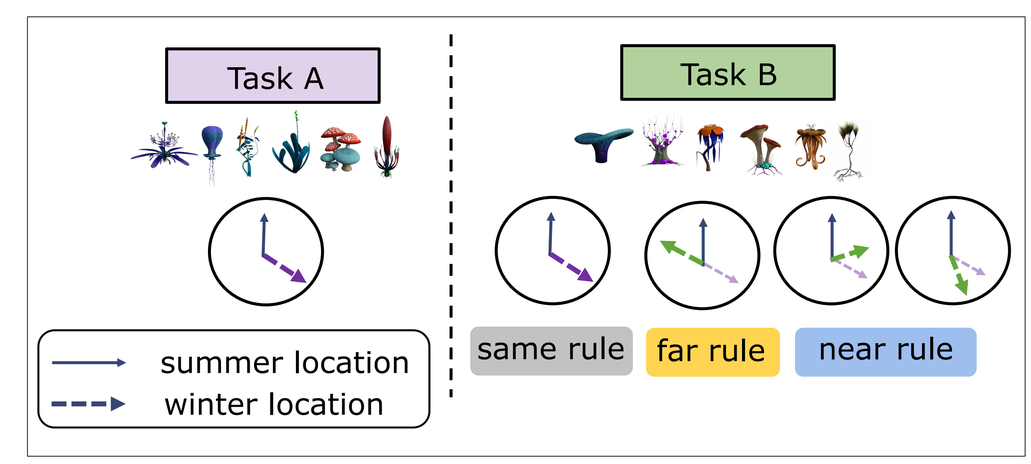
[1]Department of Experimental Psychology, University of Oxford [2]DeepMind, London

Introduction

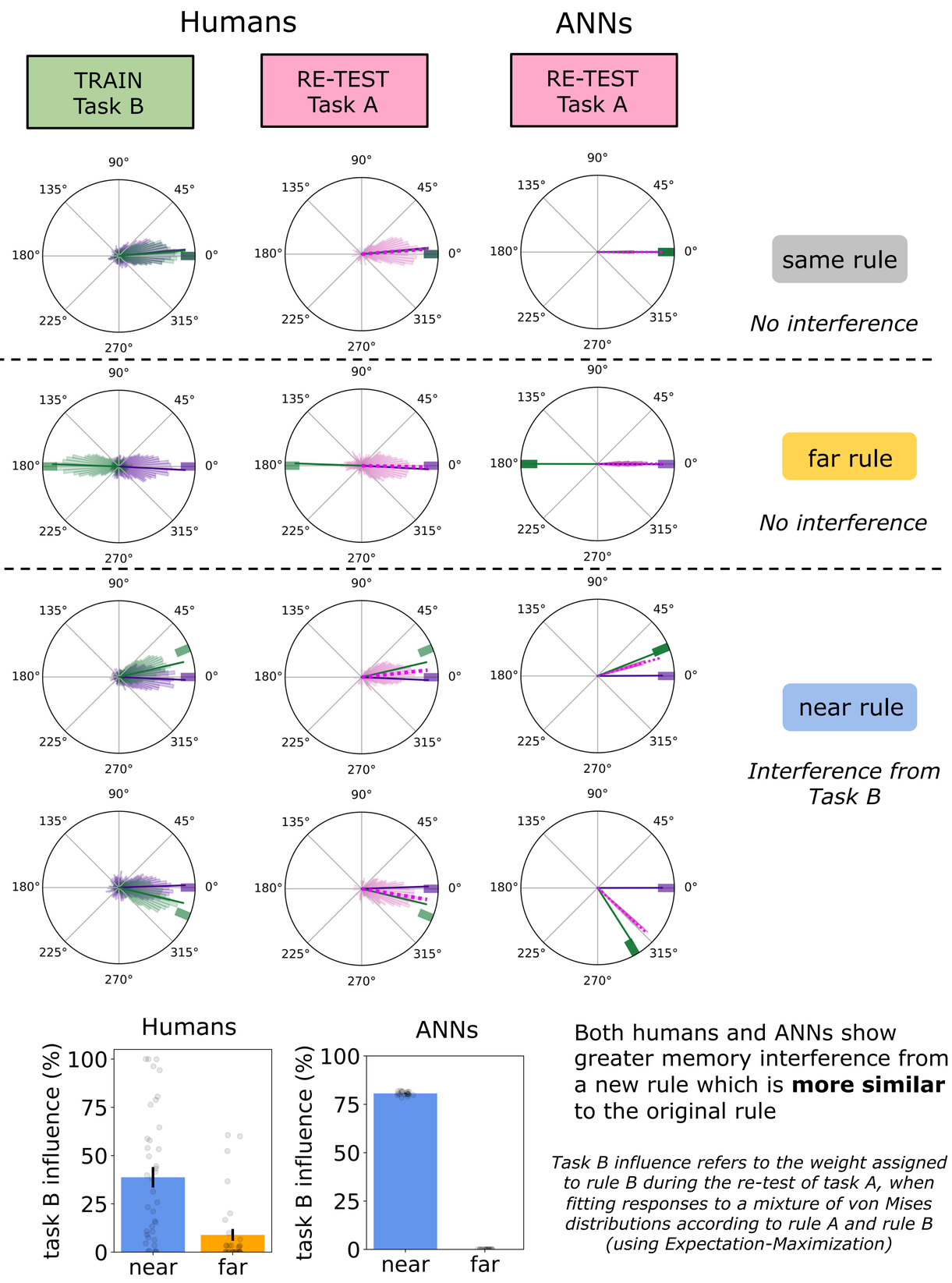
- When learning a new task, old task knowledge can incur both benefits (transfer) and costs (interference)
- Transfer and interference can **trade-off** in artificial neural networks (ANNs): re-use of existing representations leads to faster learning of new tasks alongside greater forgetting (interference) of previous tasks
- This trade-off depends on **task similarity**, with very dissimilar tasks incurring **low interference costs** since tasks are learned using new representations ("Maslow's Hammer"; Lee et al. 2022, Ramasesh et al. 2020)
- **Do humans show the same pattern when successive tasks are dissimilar: reduced transfer but lower interference?**

Paradigm





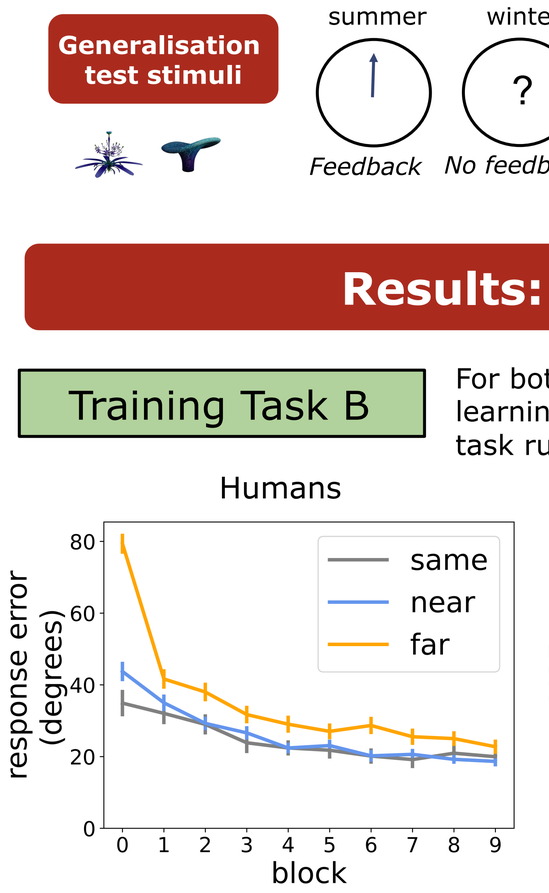
Results: Interference

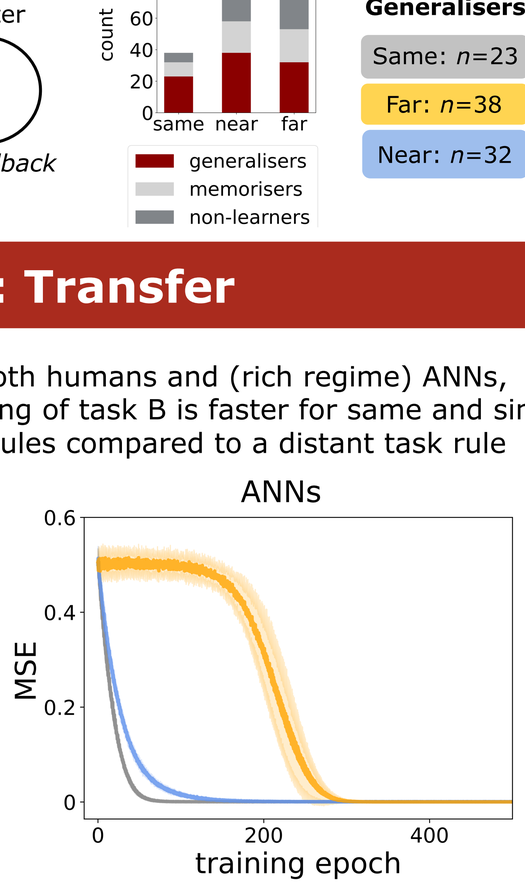


Results: Transfer

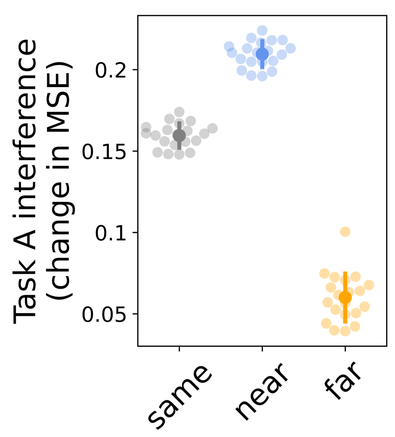
Training Task B

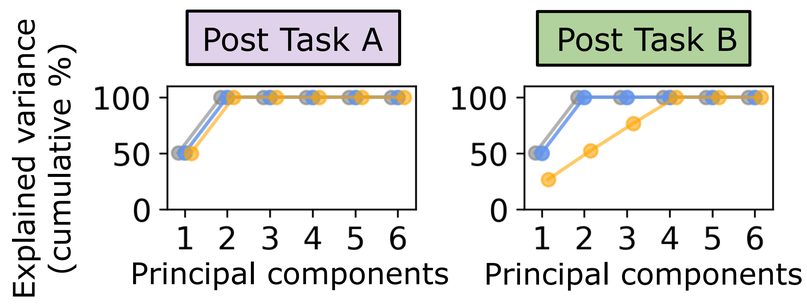
For both humans and (rich regime) ANNs, learning of task B is faster for same and similar task rules compared to a distant task rule





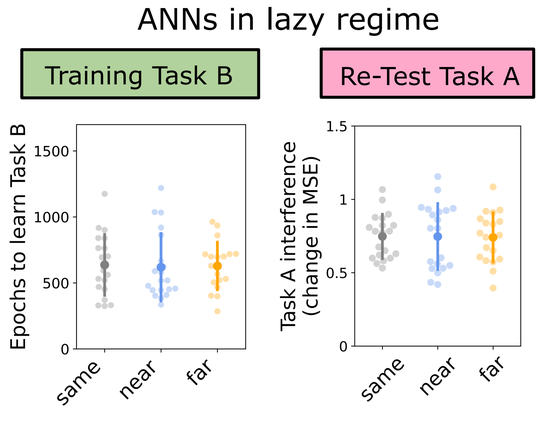
ANNs: shared or distinct task representations

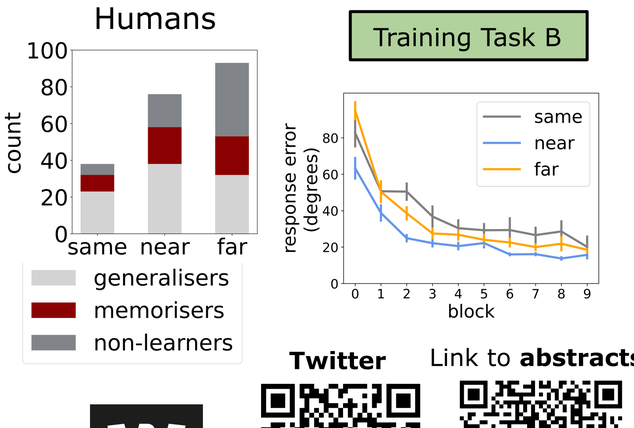




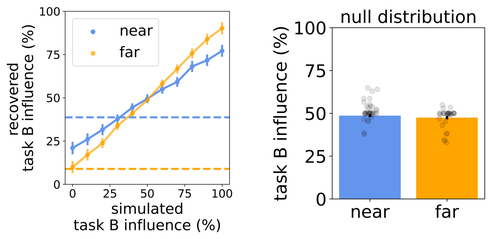
Networks trained successively on two dissimilar tasks arrive at double the number of components present in hidden layer activity, reflecting separate representations for the two tasks

Rich vs. Lazy learners





Sanity Checks



Summary

We present a paradigm where both humans and ANNs benefit from task similarity when acquiring new knowledge ('transfer') while also suffering higher interference on the original task when successive tasks are similar. In ANNs this is because the original task representation is updated through re-use for the new task, resulting in higher catastrophic forgetting.

References

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,

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

wellcome

Twitter

Link to **abstracts**

Correspondence: eleanor.holton@psy.ox.ac.uk