SPACE AS A SCAFFOLD FOR TEMPORAL GENERALISATION

Jacques Pesnot Lerousseau[1] & Christopher Summerfield[1]
[1] Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom

AGIR POUR L'AUDITION
Recherche et prévention pour la santé auditive

Paper Code & data

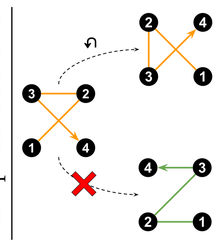
INTRODUCTION

To recognise objects and events in the natural world, humans form mental representations that are i**nvariant to transformation**. We have no trouble recognising a teapot that is moved to a new location (translated), tipped on its side (rotated) or viewed from afar (rescaled). How invariant object representations are learned is among the oldest problems in cognitive science, and has provoked diverse theories based on assembly from geometric primitives (Biederman, 1987; Marr, 1982), associative learning (Rock & DiVita, 1987; Wallis & Bülthoff, 1999) and function approximation in deep networks (Lindsay, 2021). A longstanding debate is whether invariant recognition relies on explicit encoding of the spatial structure of object features: for example, whether an invariant representation of a teapot codes the relationship among handle, body, and spout.

The idea that invariant object representations require an explicit representation of spatial structure makes two striking predictions about how humans will learn and generalise **objects defined by temporal relations among features (temporal objects, TOs)**. The first prediction is that **translation- and rotation-invariant TOs will be readily learned and generalised when sequential features are spatial locations** (e.g., x or y position) but not when they are non-spatial visual or auditory attributes such as colour or timbre. The second prediction is that **having learned to associate visual or auditory attributes with spatial locations, participants will then be able to "mentally rotate" TOs** just as they can objects defined by the spatial arrangement of their features. Thus trained, we predict that participants will successfully categorise a novel temporal sequence whose feature transitions are rotated by 90°, just as they can recognise a teapot tipped on its side. **Here, in a pre-registered study, we tested and confirmed both of these hypotheses.**

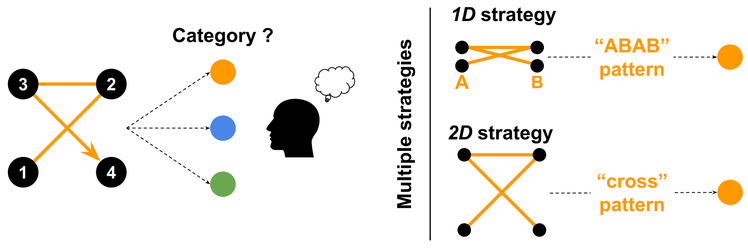
Objects with shared spatial structure

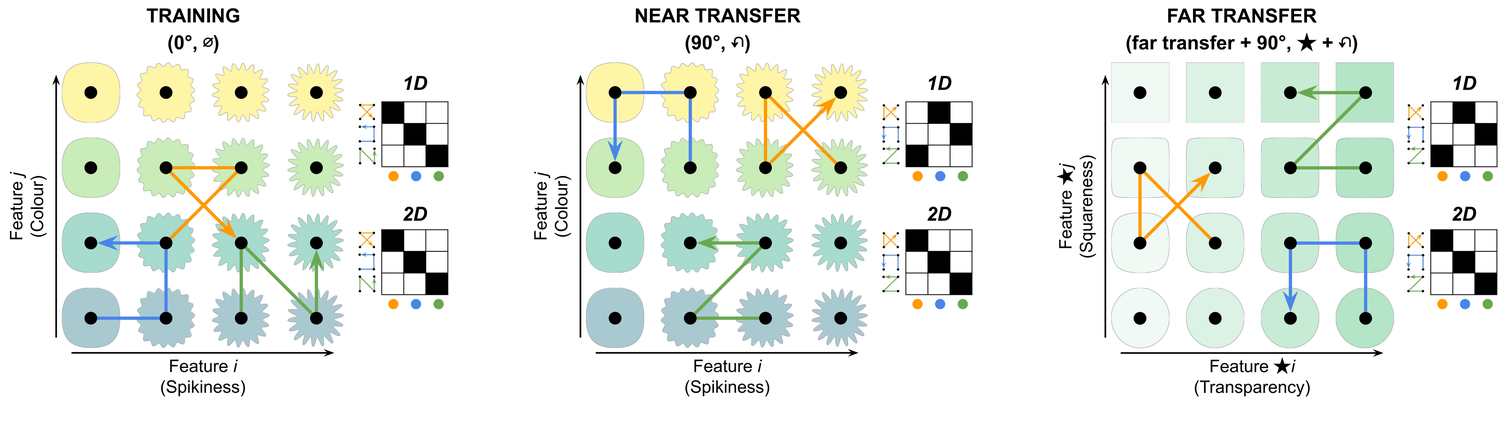
Objects with shared temporal structure



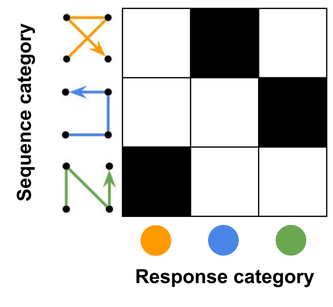
PARADIGM

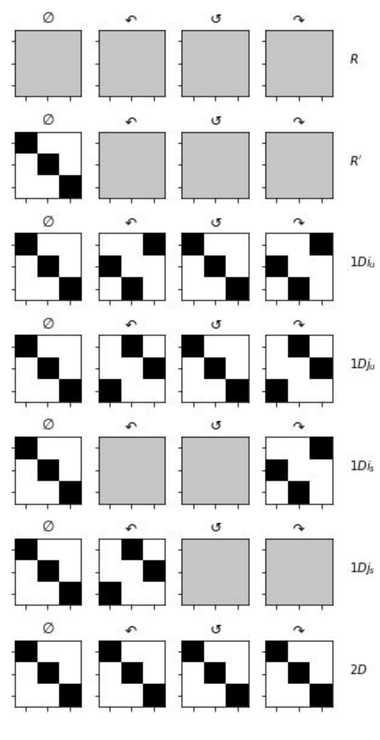
Online behavioral experiment (N = 558)
Prolific





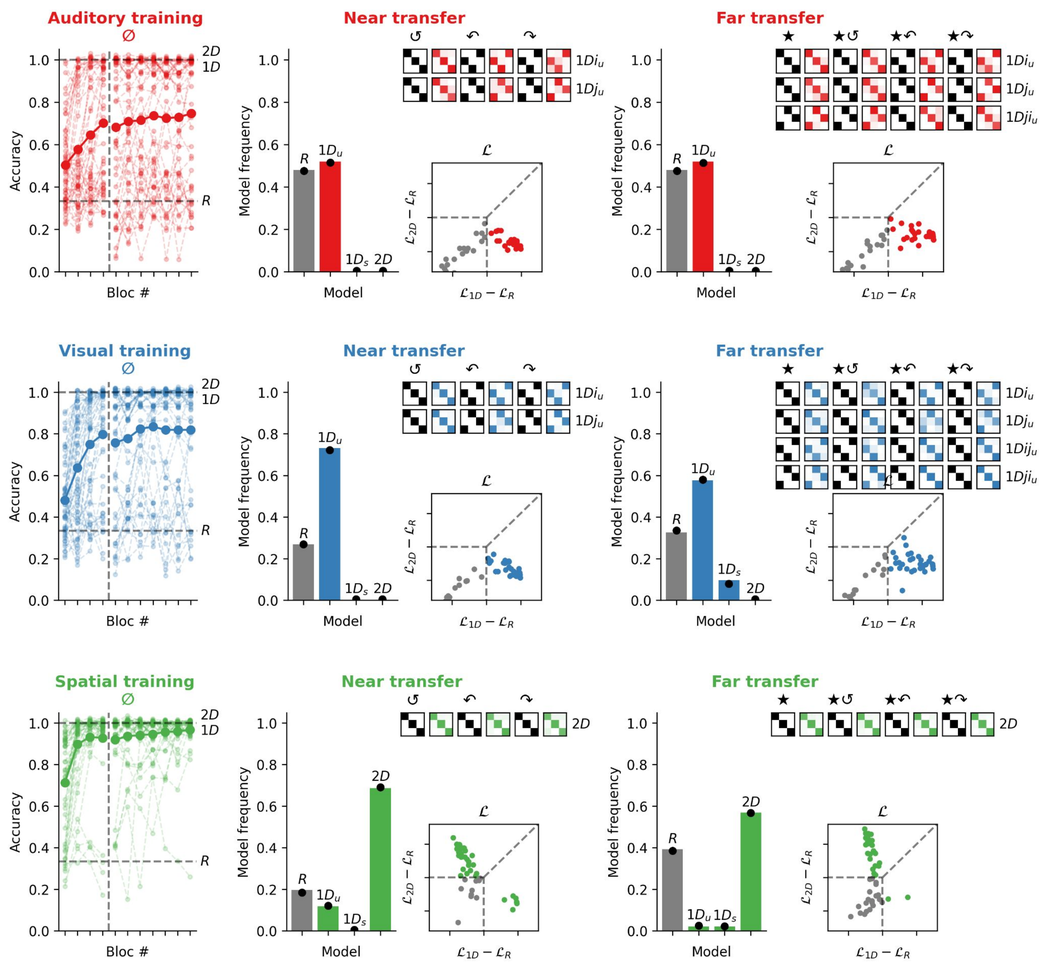
MODELS





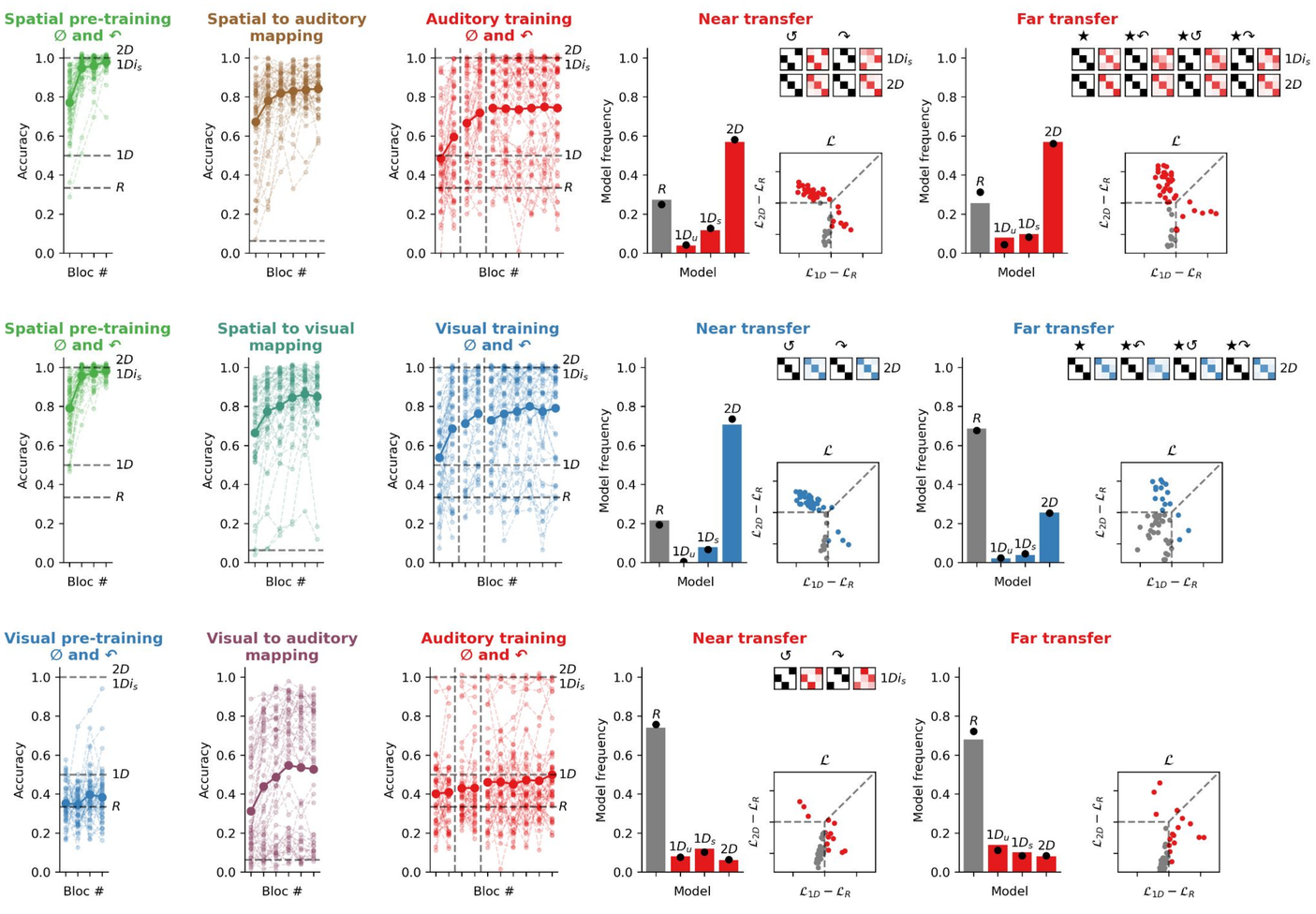
EXPERIMENT 1: MODALITY

When temporal object categories were characterised by temporal patterns in spatial location, participants learned to represent the 2D structure of the TO, and generalised readily to rotated (as well as translated) exemplars. However, when TOs were defined by patterns of auditory or visual features, participants learned mappings to each category by relying on a single feature dimension and thus failed to form rotational invariant representations.



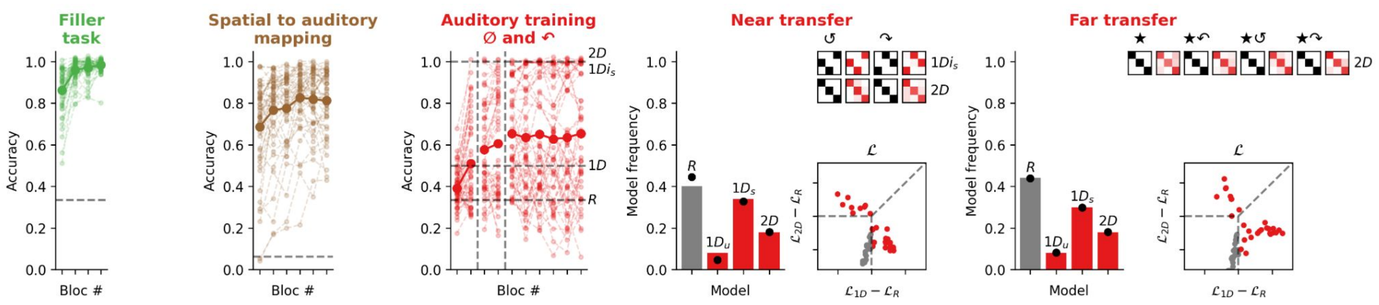
EXPERIMENT 2: SPATIAL PRETRAINING

By learning the association between auditory or visual features and a corresponding spatial location, TOs composed of exclusively auditory or visual features could now be generalised over rotations. By way of control, however, we confirmed that when the pre-training involved visual features, no such benefit occured, and participants failed to show rotation invariance. Thus, spatial pre-training provided an effective scaffold that allowed participants to learn auditory and visual temporal objects in a 2D representational format that permitted generalisation to novel rotated exemplars.



BONUS EXPERIMENT

Spatial pre-training was replaced with a duration-matched filler task. Without spatial pre-training, a sizeable proportion of participants still learned a 2D strategy in both the auditory (shown below) and visual modality.



Biederman, I. Recognition-by-components: a theory of human image understanding. Psychol. Rev. 94, 115–147 (1987).
Lindsay, G. W. Convolutional neural networks as a model of the visual system: past, present, and future. J. Cogn. Neurosci. 33, 2017–2031 (2021).
Marr, D. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. (MIT Press, 1982).
Rock, I. & DiVita, J. A case of viewer-centered object perception. Cogn. Psychol. 19, 280–293 (1987).
Wallis, G. & Bülthoff, H. Learning to recognize objects. Trends Cogn Sci (Regul Ed) 3, 22–31 (1999).

CONCLUSION

We studied whether humans can learn rotation- and translation-invariant representations of temporal objects. We found that **participants can generalise temporal object knowledge** to novel exemplars defined by **rotations of feature transition vectors**, but **only if the features were themselves physical spatial locations** (e.g., *x*, *y* position; Exp. 1) or if non-spatial attributes had **previously been mapped to a physical spatial location in a pre-training task** (Exp. 2-4). Thus, an explicit representation of space is a "scaffold" that allows generalisation in time, and an explicit representation of space is the critical factor that permits objects to be learned in an invariant fashion.