# Learning map-like neural representations of abstract concepts

Sophie M. A. Johnson[1,2,a], Sam C. Berens[1,2,b]

[1] School of Psychology, University of Sussex, Brighton, United Kingdom.
[2] Sussex Neuroscience, University of Sussex, Brighton, United Kingdom.

[a] sj483@sussex.ac.uk, https://orcid.org/0009-0009-0904-5225
[b] s.berens@sussex.ac.uk, https://orcid.org/0000-0001-8197-8745

## Preregistration statement

This is a preregistration for a study testing several fundamental hypotheses in cognitive neuroscience. Most of this preregistration was written before data collection, yet it was finalised towards the end of data collection. Some of the analyses have been informed by peeking at subsets of the final sample. Specifically, the structure of the participant-level model fitted to the training data has been heavily informed by testing analyses on a large subset of the final dataset. However, our experimental hypothesis related to this model, which will be tested with a group-level analysis, has yet to be estimated. Additionally, before this preregistration was finalised, preliminary fMRI analyses were run on a subset of the final dataset and presented during conference poster sessions. Nonetheless, the fMRI analyses outlined below were almost entirely fixed in early versions of this document before any data were collected. Further, we minimised the influence of our preliminary results on the preregistration by only running simplified versions of our planned analyses ahead of time.

## Abstract

Recent theoretical models suggest that conceptual learning is grounded in our experience with space. Just as the mammalian hippocampal system learns maps of spatial environments, it may also learn 'conceptual maps' representing inferred relations between abstract objects. Here, we test several novel hypotheses about conceptual learning that stem from these models. Across ~6 weeks, participants will learn abstract conceptual relationships between visual stimuli (a modular ring of integers). Following this, they will undergo fMRI scanning while rehearsing what they have learnt. We will use multivariate analyses to examine the structure of neural representations that have been learnt. Although the task does not explicitly involve spatial transformations, we predict that the hippocampal system will learn map-like representations of the task stimuli that encode inferred conceptual ‘distances’ between them. Mirroring what is seen in artificial neural networks, we also predict that participants will generalise what they have learnt to solve unseen problems, but only long after they have memorised explicitly trained relationships. Our study represents novel tests of key predictions made by contemporary models of conceptual learning.

# Introduction

Recent theoretical models have suggested that conceptual learning is grounded in our experience with space (Whittington et al., 2020; Momennejad, 2020). Just as we learn mental maps of familiar buildings, we may learn 'conceptual maps' that represent inferred relationships between objects (e.g., the relative “grittiness” across a set of films). Alongside these proposals, newly developed neural networks provide a powerful new framework for understanding conceptual learning (Whittington, Warren, & Behrens, 2021).

Here, we test a number of new hypotheses about conceptual knowledge acquisition that directly stem from the literature on theoretical neuroscience and machine learning. We have developed a web-based task to teach participants abstract conceptual relationships between novel visual stimuli (forming a ring of integers, $\mathrm{mod}\ 6$). Participants will spend several weeks learning these relationships and will be asked to participate in an fMRI scanning session at the end of this training period.

## Structural representation of abstract concepts

The hippocampal system is thought to facilitate the learning of structural (map-like) relationships between any set of items (e.g., locations in physical space, points in time, and animals in a social hierarchy, e.g., Constantinescu, O’Reilly & Behrens, 2016; Kumaran et al., 2016; Garvert, Dolan & Behrens, 2017). Contemporary models suggest that these structural representations are learnt by predicting the consequences of ‘movements’ in an abstract task space (e.g., “heading north from the post office will bring me to the pub”; Whittington et al., 2020; Russek et al., 2017). Critically, these models suggest that learners must make movement-based predictions even when the to-be-learnt relationships are wholly conceptual and cannot be directly experienced (e.g., “making Alice 12 units worse at chess results in Bob”).

We will test this hypothesis by training participants on a task that does not explicitly involve movement-based predictions but does imply that stimuli can be positioned in an abstract task space. Participants will see pairs of stimuli ($a + b$) that imply a third stimulus ($= c$) according to a set of mathematical rules (modular addition). They will be tasked with learning to select the implied third stimulus through trial and error. Notably, this task does not explicitly require participants to predict the consequences of ‘moving’ or ‘transitioning’ between stimuli. Nonetheless, we predict that the hippocampal system (including the hippocampus, and the entorhinal cortex) will acquire map-like representations that ‘position’ the stimuli in a ring structure according to their relative modular values.

A number of studies have suggested that the medial prefrontal cortex (MPFC) is also involved in representing structural relationships between items, particularly when those relationships are more abstract/conceptual in nature (Berens & Bird, 2022; Theves et al., 2021; Schlichting, Mumford & Preston, 2015). Given this, we predict that the medial prefrontal cortex will also acquire map-like structural representations of the stimuli.

Finally, we hypothesise that the fidelity of these structural representations (i.e., how strongly they are expressed in each individual) will predict participants’ ability to infer conceptual relationships that have not been explicitly trained. This form of structural generalisation has been studied in the context of other tasks (e.g., transitive inference; Berens & Bird, 2022) and is thought to depend on the hippocampal system and MPFC. To our knowledge, this is the first study that tests the ability of humans to generalise structural relationships in a ring of integers.

## Representing structures at different resolutions

The hippocampus is thought to represent map-like relationships between items at multiple 'resolutions' (Strange et al, 2014). Posterior regions of the hippocampus are proposed to encode fine-grained details, whereas anterior regions represent more general, longer-range relationships.

There is good evidence to support this hierarchical gradient for representations of physical space (e.g., Kjelstrup et al., 2008; Nadel, Hoscheidt, & Ryan, 2013). Additionally, there is some evidence that memory for specific vs general semantic details differently depends on posterior and anterior portions of the hippocampus, respectively (e.g., Gutchess & Schacter, 2012; Collin, Milivojevic & Doeller, 2015). However, the theory has not been explicitly tested for conceptual relationships. Moreover, it remains controversial as there are multiple, though not mutually exclusive, competing theories that can also account for some of the supporting evidence (see Poppenk, Evensmoen, Moscovitch, & Nadel, 2013).

We hypothesise that anterior portions of the hippocampus will represent stimuli in a way that downweighs small conceptual differences between stimuli relative to more posterior regions. We will test this by modelling non-linearities in the correlation between conceptual distance and representational similarity at multiple points along the hippocampal long-axis. We predict that posterior portions of the hippocampus will exhibit non-linear correlations reflecting a bias towards encoding small conceptual distances with high levels of precision. In contrast, we expect anterior portions of the hippocampus to exhibit non-linear correlations reflecting that stimuli are only represented in dissimilar ways when there is a large conceptual distance between them.

## Conceptual knowledge in sensory cortices

Visual processing has been traditionally thought of as hierarchical such that early visual regions only encode basic visual properties (e.g., lines, shapes, and colours) but not conceptual information about stimuli (Herzog & Clarke, 2014). However, visual representations learnt by convolutional neural networks reflect some degree of semantic information, even in early layers (Dharmaretnam & Fyshe, 2018). Moreover, activity in the early visual cortex is known to reflect prediction errors for high-level (e.g., whole object) predictions rather than reflecting predictions of low-level visual features (Richter et al., 2023). This suggests that the representations learnt by early visual regions are principally driven by higher-order task demands.

Despite this, there is no direct evidence that representations learnt in early visual regions actually reflect conceptual information relevant to high-level cognitive tasks. Here we will explicitly test this hypothesis using the modular addition task described above. Specifically, we predict that representations of stimuli in the primary visual cortex (V1) will be biased by the modular values of those stimuli. As such, we expect that V1 will also show evidence of map-like structural representations similar to regions in the medial temporal lobes and the MPFC.

## Delays between memorisation and generalisation

Transformers are artificial neural networks that can learn to predict the occurrence of items (e.g., words, or visual stimuli) according to complex rules pertaining to a long list of observations (e.g., passages of text, or frames of a movie, etc.; see Vaswani et al., 2017). Recently, it has been suggested adapted transformer architectures may serve as good models of the hippocampal system given that they show similarities with the Tolman-Eichenbaum Machine

(TEM), a neuroscientific model of structural learning within the hippocampal system (Whittington et al., 2021).

It is well known that transformers can learn to generalise modular arithmetic operations given a relatively small number of examples (Power et al., 2022). Surprisingly, however, there is a significant lag between when these models completely memorise the examples they are given during training, and when they begin to generalise what they have learnt to solve unseen problems (Power et al., 2022; Liu et al., 2022; Murty et al., 2023). Little is known about why this lag occurs. Nonetheless, based on the similarities between the TEM and the transformer architecture, we hypothesise that the same lag will be observable in human performance.

# Methods

## Participants and sample size

Participants will be right-handed, have either normal or corrected-to-normal vision, and have no reported history of neurological or psychiatric illness. We will aim to collect 32 usable datasets, excluding participants who must be removed from the sample during data cleaning (see below).

## Stimuli and task structure

We have produced a set of 6 abstract images referred to as 'sparks'. Sparks are composed of multiple translucent 'wisps' with varying colours and degrees of rotational symmetry rendered onto a black background (see Figure 1a). These sparks were generated using a free-to-use web application (http://weavesilk.com/) and selected from a surplus in order to maximise our ability to detect visual representations of the stimuli using fMRI.

To do this, we first quantified the visual similarity between pairs of sparks by correlating feature (activation) maps of the stimuli obtained from an early layer of a convolution neural network trained on image classification (layer 'pool1-norm1' from GoogLeNet, trained on the ImageNet dataset). We then selected stimuli such that there was a high variance and low skew of similarity scores between all pairs of images in the chosen set. This has the effect of maximising statistical power in a representational similarity analysis of visual features.

For each participant independently, sparks will be randomly assigned an integer value between 0 and 5. Training trials will present two sparks at the top of the screen, each being associated with integer values $a_i$ and $b_i$ respectively (where $i$ indexes trial number). These sparks will be followed by an abstract greyscale symbol (see Figure 1b). This array of images (referred to as the 'query') implies a modular addition between $a_i$ and $b_i$ such that the result ($c_i$) is an integer associated with a third spark or one of the sparks being queried:

$$c_i = a_i + b_i \bmod 6$$

Participants will be tasked with learning to select $c_i$ for all possible queries. In a finite field of 6 integers, there are 36 ($6^2$) ordered pairs of integers that may be subject to binary addition. We subdivide these pairs into two subsets: 'supervised pairs' and 'unsupervised pairs', containing exactly 27 and 9 members respectively (see Figure 1c). Both subsets will be presented during training, yet participants will only receive corrective feedback when responding to supervised pairs. As such, participants can only achieve above chance performance on the unsupervised pairs by generalising rules learnt from supervised training.

We further subdivide unsupervised pairs into two categories, 'commutable' and 'non-commutable' pairs (see Figure 1c). Commutable pairs may be made into a supervised pair by swapping the order of sparks in the query (e.g., if $x + y$ is unsupervised, it is also commutable when $y + x$ is supervised). In contrast, non-commutable pairs cannot be made into a supervised pair in the same way, either because both sparks in the query are identical ($a_i = b_i$), or because swapping the order of the sparks yields another unsupervised pair. We expect participants will learn to provide correct responses to commutable pairs very rapidly as the supervised training will make it apparent that the order of sparks in each query is irrelevant. Given this, we expect performance on non-commutable pairs to be our best measure of generalisation since they can only be solved by learning the modular task structure.

The allocation of pairs to subsets and categories has been designed to minimise the variance between the number of times that each integer (spark) appears in each subset/category (e.g., to avoid almost all unsupervised pairs involving one specific integer).

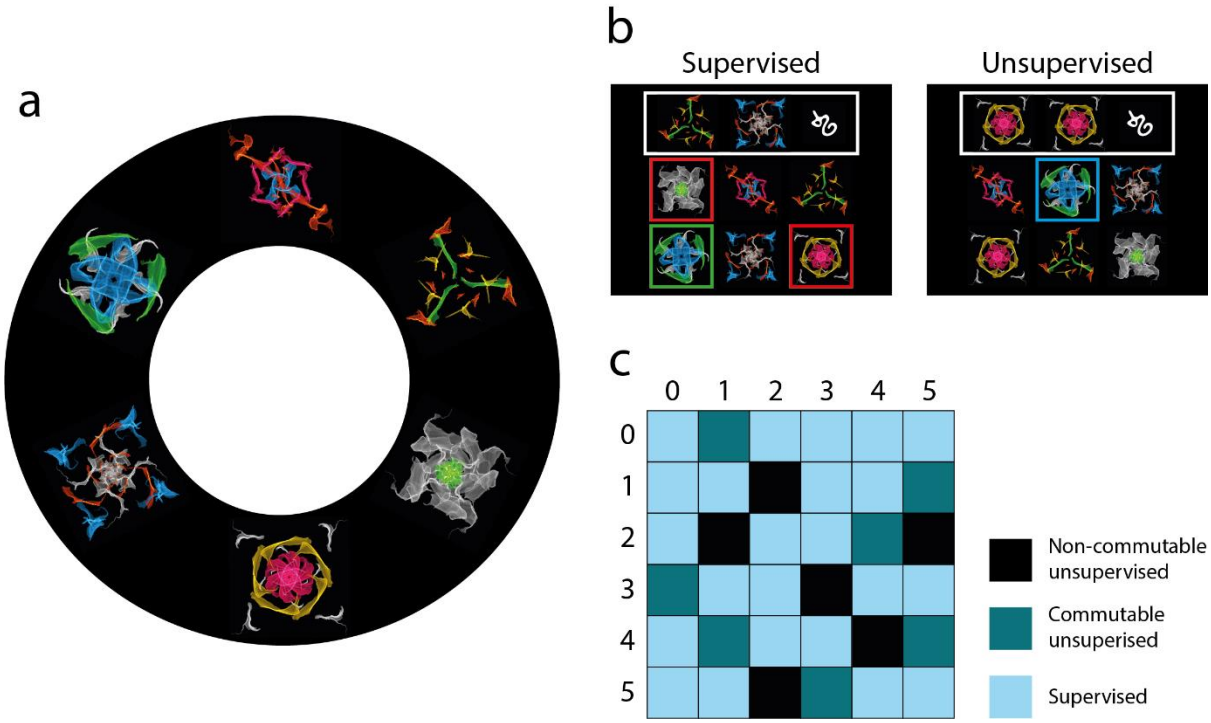


***Figure 1.*** Stimuli and task-structure. ***a)*** Stimuli (referred to as 'sparks') will be randomly assigned an integer value between 0 and 5 (independently for each participant). ***b)*** Examples of a 'supervised' and an 'unsupervised' trial from the pre-scanner training procedure. In each case, a 'query' composed of two sparks and an abstract symbol is shown at the top of the screen (white box). The query poses a modular addition between the integers associated with each spark. Below the query, the participant may select a response (the result of the addition) from a 2-by-3 array. On supervised trials, participants will receive feedback indicating whether their selection is either correct (green border) or incorrect (red border). On these trials, they will be able to continually submit responses until a correct response is provided. On unsupervised trials, participants may only submit a single response, and no corrective feedback will be given (blue border). ***c)*** A table showing the assignment of all possible modular addition queries into three mutually exclusive categories: supervised pairs, commutable unsupervised pairs, and non-commutable unsupervised pairs. Unlike non-commutable pairs, commutable pairs may be made into a supervised pair by swapping the order of sparks in the query.

## Pre-scanner training

Over the course of several weeks, participants will be invited to log into a web-based application and engage in a trial-and-error learning task that trains the modular addition operations described above (see https://github.com/Sam-Berens/Modulo_WebTask). They will not be told that the task involves a form of arithmetic, or that the sparks are associated with integer values.

Participants will be able to use any device to run the training task, including phones, tablets, and personal computers. They will be encouraged to log into the system every day and

spend at least 15 minutes on the task each time (the length of each session will be under the participant's discretion). They will also receive twice-daily push notifications from the Samply smartphone application with a polite training reminder (https://samply.uni-konstanz.de/).

On each training trial, participants will see a pair of sparks followed by an abstract greyscale symbol at the top of the screen (see Figure 1b). Below this, they will see a larger set of 6 sparks in a 2-by-3 array (the complete set of sparks being trained). The arrangement of sparks in the lower array will be randomised on each trial. Participants will be asked to predict which of the sparks in the lower array should come next in the sequence at the top of the screen. They will be able to submit their choice with a button/screen press. Participants can choose how long they take to respond, up to a maximum of two minutes. After two minutes the page will refresh, and participants will be prompted to start a new training session.

On supervised trials, participants will receive feedback indicating whether each of their responses is either correct or incorrect. In case of an incorrect response, a red border will be placed around their chosen spark (see Figure 1b). With this border remaining onscreen, participants will have repeated attempts to select the correct spark with more borders being added around any other incorrect selections. When a correct selection is made, a green border will appear around the selected spark and any existing red borders around other stimuli will be removed. This display will be shown for 4 seconds, after which the trial will end, and a new trial will begin.

On unsupervised trials, participants will not receive corrective feedback and will only have the opportunity to provide a single response on each trial. Here, a cyan border will be placed around the spark that participants first select, and this display will remain onscreen for 4 seconds before the next trial begins (see Figure 1b).

Training trials will present supervised and unsupervised pairs according to a probability distribution that makes some trials more likely to occur than others. Unsupervised pairs will be least likely to occur with an occurrence probability of $1/73$ for each pair. Supervised pairs that include the spark assigned an integer value of zero will be the next most likely to occur with an occurrence probability of $2/73$ per pair. All other supervised pairs will be most likely to occur with an occurrence probability of $3/73$ per pair. These occurrence probabilities have been chosen to minimise the total number of trials required to yield high levels of performance on all supervised pairs (pilot work indicated that participants tend to memorise pairs involving the zero-valued spark very rapidly).

Throughout training, participants will be able to view their progress in learning the supervised pairs via a dedicated webpage. This page will present a graph plotting summary statistics of their response accuracy broken down by attempt number (1st, 2nd, 3rd guesses), and training session (a continuous run of trials containing at least 9 trials). The accuracy statistics ($l_{s,k}$) for a given session ($s$) and attempt number ($k$) will be defined as follows:

$$l_{s,k} = f\left(\frac{1}{n_s}\sum_{i=1}^{n_s}\cos\left(\frac{\pi}{3}\left((\theta_{s,i,k} - t_{s,i}) \bmod 6\right)\right)\right)$$

Where: $i$ indexes the trial number within each session, $n_s$ is the number of trials within each session, $\theta_{s.i,k}$ is the modular value for a specific response, $t_{s,i}$ is the modular value of the correct (target) stimulus, and $f$ is a non-linear function that squeezes accuracy scores within a narrow interval ($[-0.135,1]$) and extends the range of accuracy statistics at higher levels of performance:

$$f(x) = \frac{\exp(2x) - 1}{\exp(2) - 1}$$

## In-scanner task

Each trial of the in-scanner task will start with the presentation of a fixation cross for 1 second. This will be followed by the presentation of two sparks ($a$ and $b$) shown one at a time for 3 seconds each, separated by a variable inter-stimulus interval (a blank screen lasting between 1 and 5 seconds ($duration \sim U(1,5)$, see Figure 2). After the presentation of spark $b$, one of three grayscale symbols will be presented for 1 second. On $88.\overline{8}\%$ of trials (8/9) the symbol will be the same as that presented during training indicating either a supervised or unsupervised query to which participants must respond. On the remaining $11.\overline{1}\%$ of trials (1/9), the symbol will be either a left (←) or a right (→) arrow, indicating that participants must respond by selecting either the first or second spark in the preceding sequence (respectively, see Figure 2). These 'n-back' trials will only follow the presentation of supervised pairs and are designed to discourage participants from computing the answer to queries before the symbol is shown.

Immediately following the symbol, a response prompt will appear for 6 seconds. During this time, participants will be able to select a spark from a 2-by-3 array. The arrangement of sparks in this array will be randomised on every trial, and a blue cursor will be drawn around one of the sparks at random. Participants will be able to scroll the cursor across all sparks in the array using an index-finger button press and make a selection using a middle-finger button press. Once a response has been entered, the cursor will turn blue to indicate that the response has been accepted and no further scrolling will be permitted. In-scanner trials will not provide corrective feedback. The next trial will begin after the 6-second response window regardless of when or whether a selection has been made.

The in-scanner task will run across several independent fMRI runs, each lasting approximately 10 minutes. We will aim to collect 5 runs of data for each participant (time permitting). During a run, 36 trials will present all supervised and unsupervised pairs (once each). Of these, 4 trials will request an n-back response following the presentation of 4 supervised pairs (randomly selected from different columns of the modular addition table). Runs will additionally include 4 null events (fixation crosses lasting 8 seconds) randomly interspersed between trials. The order of pairs and the arrangement of null events within each run has been determined by a procedure that optimises our statistical power in detecting BOLD effect of interest (see https://github.com/sj483/Modulo_ScannerTask/blob/main/OptSeq/Run01.m).

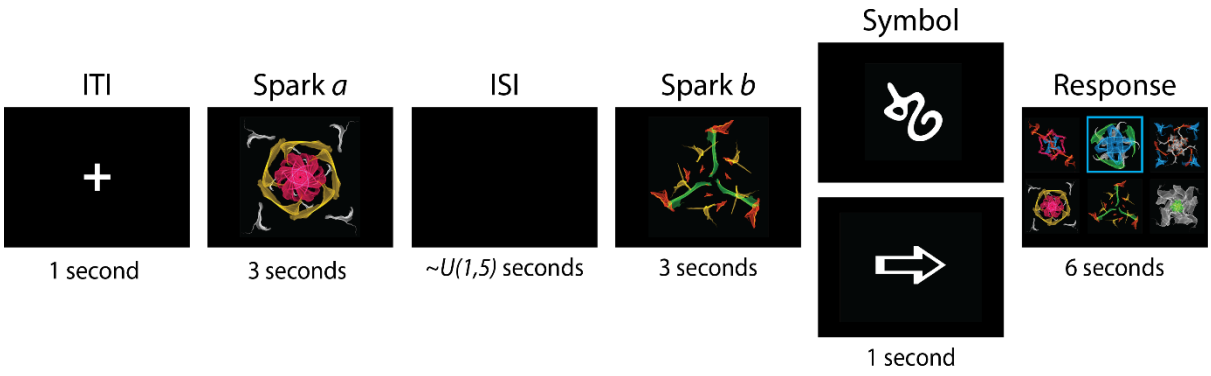


***Figure 2.*** Trial structure of the in-scanner task. *ITI*: Inter-trial interval. *ISI*: Inter-stimulus interval.

# Analyses

## Behavioural analysis of training data

### *Participant-level model*

For each participant, we will fit a hierarchical model that predicts the full sequences of observed responses across all training trials and attempts. These predictions will be based on a single continuous variable ($x_i$), intended to quantify the cumulative feedback that each participant has received across sessions. Specifically, $x_i$ will encode the mean number of times that a subset of supervised pairs has been presented prior to trial $i$. The relevant subset excludes any pair that involves a zero-valued spark (e.g., 0+0, 0+1, 1+0, …). We have defined $x_i$ in this way since learning on "zero-plus" pairs are likely to be driven by a simple heuristic ("select whatever 0 was paired with"), which will have a limited bearing on the learning of all other pairs. Given this, $x_i$ is weakly monotonic, increasing by exactly 1/14 after 14 of the 36 pairs, and remaining constant otherwise.

Within the model, values of $x_i$ will be transformed into a signed resultant vector ($r_i$) describing the expected circular distribution of responses centred on either the target spark (when $r_i > 0$) or the spark 180° from the target (when $r_i < 0$). This prediction will be based on a set of parameters ($a_j$ and $b_j$) that control the learning onset and learning rate (respectively) for pair $j \in \{0,1,2,\ldots,35\}$:

$$r_i = \tanh\left(b_j \ln\left(1 + \exp\left(x_i - a_j\right)\right)\right)$$

The resultant vector will then be transformed into a von Mises concentration parameter ($\kappa_i$) via an approximation provided by Fisher (1993):

$$\kappa_i = \begin{cases} 2r_i + r_i^3 + \dfrac{5r_i^5}{6}, & |r_i| < 0.53 \\ -0.4 + 1.39r_i + \dfrac{0.43}{1 - r_i}, & 0.53 \le |r_i| < 0.85 \\ \dfrac{1}{3r_i - 4r_i^2 + r_i^3}, & |r_i| \le 0.85 \end{cases}$$

This concentration parameter will then be used to compute a discrete probability distribution ($\hat{p}_i$) over all possible responses for the first attempt on trial $i$:

$$\hat{p}_i(y) = \exp\left(\kappa_i \cos\left(\frac{\pi}{3}(y - t_i)\right)\right) c_i$$

Where, $y$ is the integer value of each response possibility, $t_i$ is the integer value of the target, $y, t_i \in \{0,1,\ldots,5\}$, and $c_i$ is chosen to ensure that the sum of all response probabilities on a given trial equals one. Given this per-trial probability distribution, we will then compute the expected probability of any sequence of responses ($\hat{q}_i$) on trial $i$ assuming a sequential selection process without replacement (as enforced by the training task):

$$\hat{q}_i = \prod_{m=1}^{k_i} \frac{\hat{p}_i(y_{i,m})}{1 - \sum_{n=1}^{m-1} \hat{p}_i(y_{i,n})}$$

Where, $m$ and $n$ index the attempt number and $k_i$ is the total number of attempts made on trial $i$. Using these per-trail probabilities, we will then define a log-likelihood statistic over all trails to guide model fitting:

$$\ell = \sum_i \ln(\hat{q}_i)$$

This model will be estimated in CmdStan using the No U-Term Sampler (v2.36.0; Stan Development Team, 2024). This involves drawing samples from a posterior probability distribution of model parameters and using these samples to estimate expected values (means), and credible intervals.

A set of hierarchical priors will be used to regularise the learning onset and learning rate parameters, $a_j$ and $b_j$. Specifically, we classify each of the 36 pairs into four mutually exclusive categories whose members are assumed to be well-described by similar parameter values: *1)* supervised pairs that do not include a zero-valued spark, *2)* all pairs involving a zero-valued spark (supervised or unsupervised), *3)* unsupervised pairs that are commutable, and *4)* unsupervised pairs that are non-commutable. The $a_j$ and $b_j$ parameters will also be centred and transformed to allow for more efficient sampling and ensure that they remain within valid bounds:

$$a_j = S(\alpha_{0,g} + \alpha_{1,g} u_j) \times \max(x)$$

$$b_j = \beta_{0,g} + \beta_{1,g} v_j$$

Where, $S()$ is the standard sigmoid function, $g$ indexes each of the above categories ($g \in \{0,1,2,3\}$), $\alpha_{0,g}$ and $\beta_{0,g}$ are category-level parameters encoding the mean learning onset and rate for each group (respectively), and $\alpha_{1,g}$ and $\beta_{1,g}$ are category-level parameters encoding the standard deviation learning onset and rate within each group. We will assume that the pair-specific free parameters $u_j$ and $v_j$ to be independent and normally distributed:

$$u_j \sim \mathcal{N}(0,1)$$

$$v_j \sim \mathcal{N}(0,1)$$

Finally, we place the following hyper-priors on the category-level parameters:

$$\alpha_{0,g} \sim Logistic(0,1)$$

$$\beta_{0,g} \sim \mathcal{N}(0,1)$$

$$\alpha_{1,g} \sim \mathcal{N}(1, 0.05^2)\ truncated\ to\ [0,\infty)$$

$$\beta_{1,g} \sim \mathcal{N}(0.1, 0.005^2)\ truncated\ to\ [0,\infty)$$

Stan code implementing this model is available at: https://github.com/Sam-Berens/Modulo_Analysis/blob/master/A00/hspvm.stan

### *Group-level model*

To test our hypothesis that there is a substantial the delay between memorisation and generalisation, posterior mean estimates from the participant-level model will be analysed within a group-level mixed-effects regression. To do this, we will first define a new outcome variable ($\chi$) that encodes the average number of supervised trails (i.e., values of $x_i$ define above) required before participants: *a)* memorise the non-zero-plus supervised pairs, and *b)* start to show performance increases on all other pairs:

$$\chi_{j,s} = \begin{cases} a_{j,s} + \tau_{j,s}, & j \in Sup \\ a_{j,s}, & \text{otherwise} \end{cases}$$

Where, $a_{j,s}$ is the learning onset parameter (estimated as above) of a given pair (indexed by $j$) and participant (indexed by $s$), $Sup$ is the set of supervised pairs excluding zero-plus queries, and $\tau_{j,s}$ is a model-estimated learning period that quantifies the delay between learning onset and memorisation for a specific pair, dependent on that pair's learning rate ($b_{j,s}$):

$$\tau_{j,s} = \ln\left(\exp\left(\frac{\tanh^{-1}(r_{crit})}{b_{j,s}}\right) - 1\right)$$

$r_{crit}$ is a threshold that determines what level of behavioural performance defines the point of memorisation. Here, we set $r_{crit} = 0.95$ as this approximately corresponds to a predicted probability of a correct response of $0.99$ on the first attempt of a trial. Next, we will enter all values of $\chi_{j,s}$ into group-level mixed effects regression using the following model formula (specified using Wilkinson notation):

$$\chi \sim 1 + learnt * (zero + commute + noncom) + (1|pair) + (1|subject)$$

$zero$, $commute$, and $noncom$ are binary-coded fixed effects predictors that classify all pairs into four mutually exclusive types: *1)* supervised pairs that do not include a zero-valued spark (represented by the intercept), *2)* all pairs involving a zero-valued spark (supervised or unsupervised), *3)* unsupervised pairs that are commutable, and *4)* unsupervised pairs that are non-commutable. Additionally, $learnt$ is a binary fixed-effect predictor that encodes whether performance for each pair being modelled is more likely than not to result in a correct response on the first attempt of the final training trial (expected when $r_I > 0.5$, with $I$ indexing the final trial):

$$learnt = \begin{cases} 1, & Pr(r_I > 0.5) > 0.5 \\ 0, & \text{otherwise} \end{cases}$$

Finally, the model will include random intercepts for each pair and subject. It will use an identity link function and will be estimated via maximum likelihood in MATLAB.

Our hypothesis regarding the delay between memorisation and generalisation implies that the $learnt{:}noncom$ interaction term will be significantly greater than zero (i.e., supervised learning offsets are larger learning onsets for non-commutable unsupervised trials, specifically when those pairs are well-learnt by the end of training.

## MRI Preprocessing

Image pre-processing will be performed in SPM25 (www.fil.ion.ucl.ac.uk/spm). This will involve spatially realigning all EPI volumes to the first image in the time series. At the same time, images will be corrected for geometric distortions caused by field inhomogeneities (as well as the interaction between motion and such distortions) using the Realign and Unwarp algorithms in SPM (Andersson et al., 2001; Hutton et al., 2002). Following this, EPI volumes will be corrected for inter-slice acquisition delay via sinc-interpolation.

We will estimate multivariate BOLD responses to each spark in both the '$a$' and '$b$' positions of the in-scanner queries (resulting in 12 BOLD patterns of interest). These estimates will be derived from a set of first-level general linear models (GLMs) of the unsmoothed EPI data in native space using the least-squares-separate method (Mumford et al., 2012). Specifically, a unique GLM will be constructed for each spark in each position such that one event regressor models the effect of that spark while another regressor accounts for all other sparks in the time series. As such, one beta estimate from every model will encode the response for a particular spark in a particular position.

The GLMs will also include the following nuisance regressors: 1 decision/response period regressor, 6 affine motion parameters, their first-order derivatives, a Fourier basis set implementing a 1/128 Hz high-pass filter, and an intercept term. A variable number of "one-hot" binary predictors will censor out volumes where the mean framewise displacement is larger than 0.5 mm. Additionally, entire runs will be excluded if more than 20% of volumes are censored out based on this threshold.

## ROI selection

We will test our a priori hypotheses in five regions of interest (ROIs): *1)* the anterior and *2)* posterior hippocampus, *3)* the entorhinal cortex, *4)* medial prefrontal cortex, and *5)* the primary visual cortex (V1). Effects of interest will be computed separately for regions the left and right hemispheres and analysed within the same group-level regression model. All ROIs will be defined by binary masks warped to native space for each participant. The masks will be obtained by transforming group-level masks in MNI space using DARTEL's inverse warp utility in SPM25.

For the hippocampus, we will use a set of head, body, and tail masks provided by Ritchey et al. (2015). Here, we define the anterior hippocampus as the head of the hippocampus and the posterior region as the union of the body and tail. The entorhinal masks will be derived from the maximum probability tissue labels provided by Neuromorphometrics Inc.

Two separate masks corresponding to the left and right MPFC will be defined from a parcellation that divided the cortex into 200 clusters based on 17 resting-state networks identified by Schaefer et al (2018; region numbers: 82 and 187). This parcellation will also be used to define the left and right early visual cortex ROI (parcels 1 and 101 respectively).

## Representational similarity analyses

Multivariate BOLD patterns for each spark in the '$a$' and '$b$' positions will be estimated (as above) and extracted for each ROI. The Pearson correlation between these patterns will be taken as a measure of representational similarity (denoted $r_{x,y}$ for sparks that have integer values $x$ and $y$). The observed similarity between each pair of sparks will then be compared to a predicted similarity ($\widehat{r_{x,y}}$) structure, defined as follows:

$$\widehat{r_{x,y}} = 1 - \frac{\min\big((x-y) \bmod 6, (y-x) \bmod 6\big)}{3}$$

We will qualify the strength of the predicted representations in a given ROI as the Fisher-transformed Persons correlation ($z$) between predicted and observed similarity statistics:

$$z = \tanh^{-1}\big(\mathrm{corr}(r, \hat{r})\big)$$

Two Fisher-*z* statistics will be computed for each participant, ROI, and hemisphere: one of these will express the strength of modular representations for stimuli presented in the same position (i.e., comparisons between sparks presented in either the '$a$' or '$b$' positions). The other *z* statistic will express the strength of the modular representations for stimuli presented in different positions (i.e., comparisons between one spark presented in the '$a$' position, and another presented in the '$b$' position). Notably, both of these statistics will not be weighted by the similarity between sparks of the same modular value (i.e., when both sparks are visually identical). This ensures that our measure of the strength of conceptual representations was not confounded by perceptual similarity.

For a given ROI, we will then model the magnitude of Fisher-*z* statistics for both same- and different-position comparisons across all participants in a mixed-effects model:

$$z \sim 1 + colocation + hemisphere + p\_noncom + (1|subject)$$

Where, $colocation$ is a bipolar predictor coding *z*-statistics from across- and within-position comparisons as $-1$ and $1$ (respectively), and $hemisphere$ is a bipolar predictor coding observations from the left and right sides of the brain as -1 and $1$ (respectively). $p\_noncom$ is a continuous mean-centred predictor encoding the summed probability of a correct response for all non-commutable pairs at the end of training, as estimated by the participant-level analysis described above. The model will include random intercepts for each subject. Random slopes for $colocation$ and $hemisphere$ will be added on the basis of whether they improve the model fit as measured by the BIC statistic. The model will use an identity link function and will be estimated via maximum likelihood in MATLAB.

Our hypothesis regarding the presence of conceptual representations in the hippocampal system, MPFC, and visual cortices dictate that average Fisher-*z* scores will be greater than zero. This will be tested in the above model via a one-sample *t*-tests on the intercept term:

$$H_2\colon\ \beta_0 > 0;$$

Our hypothesis that the strength of structural representations will be predicted by generalisation performance implies that the model coefficient expressing the size of this effect should be greater than zero. This will be tested via a one-sample *t*-tests:

$$H_3\colon \beta_{p\_noncom} > 0;$$

Given that we are testing these hypotheses in multiple ROIs, we will use the Holm-Bonferroni method to protect the Type 1 error rate within two independent families (treating $H_2$ and $H_3$ separately). One family will include all ROIs in the medial temporal and frontal lobes (i.e., the anterior & posterior hippocampus, the entorhinal cortex, and the MPFC; 4 ROIs in total). A second family will include the early visual cortex ROIs.

Our final hypothesis dictates that anterior portions of the hippocampus will downweigh large conceptual differences between stimuli relative to posterior regions. We predict this will be detectable by analysing how the representational similarity between pairs of sparks is related to their modular distance. To do this within a given ROI or searchlight, we will first compute the arithmetic mean of all similarity scores ($r_{x,y}$) that have the same predicted value according to our structural hypothesis (i.e., a specific $\widehat{r_{x,y}}$, excluding where this value equals 1, including both within and across positions comparisons). This will yield 3 mean similarity scores corresponding to pairs of sparks with modular distances of 1, 2, and 3 denoted $\rho_1$, $\rho_2$, and $\rho_3$ (respectively). We will then model these mean similarity scores as a function of modular distance:

$$\rho_i = b_0 + b_1\left(1 - \frac{i}{3}\right)^q$$

Where $i$ is the modular distance between pairs of sparks, and $b_0$, $b_1$ and $q$ are free parameters estimated for the data. Our structural hypothesis implies that $\rho_1 > \rho_2 > \rho_3$. When this inequality holds, the model will perfectly fit mean similarity scores and $q$ will be a real number in the interval $[0, \infty]$. In such cases, $q$ encodes non-linearities in the relationship between modular distance representational similarity. Specifically, when $q = 1$, there is a strictly linear relationship between modular distance and representational similarity. When $q > 1$,

representational similarity changes most rapidly across relatively small modular distances (e.g., between $\rho_1$ and $\rho_2$). When $q < 1$, representational similarity changes most rapidly across relatively large modular distance (e.g., between $\rho_2$ and $\rho_3$).

Given this, we will test whether portions of the hippocampal long-axis differentially encode large versus small modular distances by modelling how fitted values of $q$ vary across the long-axis. To do this, we will estimate multiple values of $q$ in a searchlight analysis performed in native space for each participant (searchlight radius: 3 voxels). Values of $q$ will only be accepted if the $\rho_1 > \rho_2 > \rho_3$ inequality holds and searchlights will only be considered if their central voxel is included in our hippocampal ROIs (including the head, body and tail). The resulting image of $q$ values will then be warped to MNI space. Next, we will compute a Fisher-transformed correlation statistic for each participant that reflects the strength of the association between log-transformed values of $q$ from each searchlight ($q_i$), and long-axis position, $y_i$ (the anterior-posterior MNI coordinate):

$$\omega = \tanh^{-1}\big(\mathrm{corr}(y_i, \ln(q_i))\big)$$

Our hypothesis dictates that values of $\omega$ will be below zero and this will be tested across participants via a one-sample *t*-test against zero:

$$H_4: \frac{1}{N}\sum_{k=1}^{N} \omega_k < 0;$$

Where, $\omega_k$ is the $\omega$ statistic for the $k^{\text{th}}$ participant.

# References


Andersson, J. L., Hutton, C., Ashburner, J., Turner, R., & Friston, K. (2001). Modeling geometric deformations in EPI time series. *Neuroimage, 13*(5), 903-919. https://doi.org/10.1006/nimg.2001.0746

Berens, S. C., & Bird, C. M. (2022). Hippocampal and medial prefrontal cortices encode structural task representations following progressive and interleaved training schedules. *PLoS Computational Biology, 18*(10), e1010566. https://doi.org/10.1371/journal.pcbi.1010566

Chersi, F., & Burgess, N. (2015). The cognitive architecture of spatial navigation: hippocampal and striatal contributions. *Neuron, 88*(1), 64-77. https://doi.org/10.1016/j.neuron.2015.09.021

Collin, S. H., Milivojevic, B., & Doeller, C. F. (2015). Memory hierarchies map onto the hippocampal long axis in humans. *Nature neuroscience, 18*(11), 1562-1564. https://doi.org/10.1038/nn.4138

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science, 352*(6292), 1464-1468. https://doi.org/10.1126/science.aaf0941

Dharmaretnam, D., & Fyshe, A. (2018). The emergence of semantics in neural network representations of visual information. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2, 776-780. https://doi.org/10.18653/v1/N18-2122

Fisher, N. I. (1993). *Statistical analysis of circular data*. Cambridge University Press (Cambridge, UK). https://doi:10.1017/CBO9780511564345.

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife, 6*, e17086. https://doi.org/10.7554/eLife.17086

Gutchess, A. H., & Schacter, D. L. (2012). The neural correlates of gist-based true and false recognition. *Neuroimage, 59*(4), 3418-3426. https://doi.org/10.1016/j.neuroimage.2011.11.078

Herzog, M. H., & Clarke, A. M. (2014). Why vision is not both hierarchical and feedforward. *Frontiers in computational neuroscience, 8*, 135. https://doi.org/10.3389/fncom.2014.00135

Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., & Turner, R. (2002). Image distortion correction in fMRI: a quantitative evaluation. *Neuroimage, 16*(1), 217-240. https://doi.org/10.1006/nimg.2001.1054

Kjelstrup, K. B., Solstad, T., Brun, V. H., Hafting, T., Leutgeb, S., Witter, M. P., ... & Moser, M. B. (2008). Finite scale of spatial representation in the hippocampus. *Science, 321*(5885), 140-143. https://doi.org/10.1126/science.1157086

Kumaran, D. (2012). What representations and computations underpin the contribution of the hippocampus to generalization and inference?. *Frontiers in Human Neuroscience, 6*, 157. https://doi.org/10.3389/fnhum.2012.00157

Kumaran, D., Banino, A., Blundell, C., Hassabis, D., & Dayan, P. (2016). Computations underlying social hierarchy learning: distinct neural mechanisms for updating and representing self-relevant information. *Neuron, 92*(5), 1135-1147. https://doi.org/10.1016/j.neuron.2016.10.052

Liu, Z., Kitouni, O., Nolte, N. S., Michaud, E., Tegmark, M., & Williams, M. (2022). Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems, 35*, 34651-34663. https://doi.org/10.48550/arXiv.2205.10343

Momennejad, I. (2020). Learning structures: predictive representations, replay, and generalization. *Current Opinion in Behavioral Sciences, 32*, 155-166. https://doi.org/10.1016/j.cobeha.2020.02.017

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage, 59*(3), 2636-2643. https://doi.org/10.1016/j.neuroimage.2011.08.076

Murty, S., Sharma, P., Andreas, J., & Manning, C. D. (2023). Grokking of hierarchical structure in vanilla transformers. *arXiv*, arXiv:2305.18741. https://doi.org/10.48550/arXiv.2305.18741

Nadel, L., Hoscheidt, S., & Ryan, L. R. (2013). Spatial cognition and the hippocampus: the anterior–posterior axis. *Journal of Cognitive Neuroscience, 25*(1), 22-28. https://doi.org/10.1162/jocn_a_00313

Poppenk, J., Evensmoen, H. R., Moscovitch, M., & Nadel, L. (2013). Long-axis specialization of the human hippocampus. *Trends in Cognitive Sciences, 17*(5), 230-240. https://doi.org/10.1016/j.tics.2013.03.005

Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv*, arXiv:2201.02177. https://doi.org/10.48550/arXiv.2201.02177

Ritchey, M., Montchal, M. E., Yonelinas, A. P., & Ranganath, C. (2015). Delay-dependent contributions of medial temporal lobe regions to episodic memory retrieval. *eLife, 4*, e05025. https://doi.org/10.7554/eLife.05025

Richter, D., Kietzmann, T. C., & Lange, F. P. de. (2023). High-level prediction errors in low-level visual cortex. *bioRxiv*. https://doi.org/10.1101/2023.08.21.554095

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Computational Biology, 13*(9), e1005768. https://doi.org/10.1371/journal.pcbi.1005768

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X. N., Holmes, A. J., ... & Yeo, B. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral cortex, 28*(9), 3095-3114. https://doi.org/10.1093/cercor/bhx179

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature communications, 6*(1), 8151. https://doi.org/10.1038/ncomms9151

Stan Development Team (2024). Stan Reference Manual, v2.36.0. https://mc-stan.org

Strange, B. A., Witter, M. P., Lein, E. S., & Moser, E. I. (2014). Functional organization of the hippocampal longitudinal axis. *Nature reviews neuroscience, 15*(10), 655-669. https://doi.org/10.1038/nrn3785

Theves, S., Fernandez, G., & Doeller, C. F. (2019). The Hippocampus Encodes Distances in Multidimensional Feature Space. *Current Biology, 29*(7), 1226-1231.e3. https://doi.org/10.1016/j.cub.2019.02.035

Theves, S., Neville, D. A., Fernández, G., & Doeller, C. F. (2021). Learning and representation of hierarchical concepts in hippocampus and prefrontal cortex. *Journal of Neuroscience, 41*(36), 7675-7686. https://doi.org/10.1523/jneurosci.0657-21.2021

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*. https://doi.org/10.48550/arXiv.1706.03762

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2020). The Tolman-Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell, 183*(5), 1249-1263. https://doi.org/10.1016/j.cell.2020.10.024

Whittington, J. C., Warren, J., & Behrens, T. E. (2021). Relating transformers to models and neural representations of the hippocampal formation. *arXiv*, arXiv:2112.04035. https://doi.org/10.48550/arXiv.2112.04035