A neural mechanism for compositional generalization of structure in humans

Lennart Luettgau[1,2], Nan Chen[1,2], Tore Erdmann[1,3], Sebastijan Veselic[2,4], Rani Moran[1,2,5], Zeb Kurth-Nelson[1,6*], Raymond J. Dolan[1,2*]

[1] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, United Kingdom
[2] Wellcome Centre for Human Neuroimaging, University College London, United Kingdom
[3] Division of Psychiatry, University College London, United Kingdom
[4] Clinical and Movement Neurosciences, Department of Motor Neuroscience, University College London, United Kingdom
[5] School of Biological and Behavioural Sciences, Queen Mary University of London, United Kingdom
[6] Google DeepMind, London, United Kingdom
[*] equal contribution

Abstract

An exceptional human ability to adapt to the dynamics of novel environments relies on abstracting and generalizing past experiences. While previous research has examined how humans generalize isolated sequential processes, we know little concerning the neural mechanisms that enable adaptation to the more complex dynamics that govern everyday experience. Here, we deployed a novel sequence learning task based on graph factorization, coupled with simultaneous magnetoencephalography (MEG) recordings, to ask whether reuse of experiential “building blocks” provides an abstract structural scaffolding that enables inference and generalization. We provide behavioral evidence that participants decomposed task experience into subprocesses, abstracted dynamical subprocess structures away from sensory specifics, and transferred these to a new task environment. Neurally we show this transfer is underpinned by a representational alignment of abstract subprocesses across task phases, where this included enhanced neural similarity among stimuli that adhered to the same subprocess, a temporally evolving mapping between predictive representations of subprocesses and a generalization of the precise dynamical roles that stimuli occupy within graph structures. Crucially, decoding strength for dynamical role representations predicted behavioral success in transfer of subprocess knowledge, consistent with a role in supporting behavioral adaptation in new environments. We propose a structural scaffolding mechanism enables compositional generalization of dynamical subprocesses that facilitate efficient adaptation within new contexts.

Introduction

Among the defining features of human cognition is an exceptional ability to adapt rapidly to novel contexts. One set of proposals suggests this relies on abstraction and generalization of past experiences (Allen et al., 2020; Dekker et al., 2022; Kumar et al., 2022; Lake et al., 2015, 2017; Lehnert et al., 2020; Tsividis et al., 2021), enabling a transfer of past experiential knowledge to entirely new situations (Behrens et al., 2018; Lake et al., 2017; Mark et al., 2020; Shepard, 1987). Much previous research has been devoted to understanding how humans detect (Turk-Browne et al., 2008), generalize (Reber, 1967) and neurally represent a single dynamical, and sequentially, unfolding process (Garvert et al., 2017; Henin et al., 2021; Schapiro et al., 2013; Sherman et al., 2020) or do so from piecemeal presentation of a graph structure (Rmus et al., 2022). Here, human neural evidence suggests these processes relate to representation of experiential elements in the hippocampal-entorhinal system (Garvert et al., 2023; Zheng et al., 2024). However, despite this progress, we know much less regarding how humans abstract from the more complex dynamics of everyday experience.

Computational modeling on artificial neural networks, trained to perform multiple tasks (Johnston & Fusi, 2023; Yang et al., 2019), as well as cross-species neural evidence, support a notion that low dimensional neural coding of abstract, disentangled, task variables within structures including hippocampal-entorhinal system (Baram et al., 2021; Bernardi et al., 2020; Courellis et al., 2024), frontal cortex (Baram et al., 2021; Bernardi et al., 2020; Flesch et al., 2022, 2023; Zhou et al., 2021) and posterior parietal cortex (Flesch et al., 2022, 2023) enables a linear readout of relevant abstracted task dimensions (Bernardi et al., 2020; Courellis et al., 2024) to support inference and generalization of knowledge. This proposed mechanism may allow for computationally low demanding reuse or “recycling” of learnt representations in novel contexts.

In these studies, only a single dynamical process/structure needed to be considered and generalized to new situations. However, our everyday world experience is often the product of more complex environmental dynamics, comprising a *multitude* of simultaneously evolving structural subprocesses. Indeed, despite the idiosyncrasies and complexity of every novel situation in life, for the greater part new experiences share structural similarities with previous experiences. For instance, imagine learning to cook a new dinner recipe from a TV show while also listening to the radio. The sequence of experienced events might involve increasing the heat, adding new ingredients, then reducing the heat and stirring periodically – all while listening to music, weather updates, and traffic reports in turn. Both activities – cooking and listening to the radio – combine to shape the overall experience of the evening. In theory, we could either learn the entire sequence of cooking and radio programs together, or separately

understand the process of cooking and the structure of the radio program. The latter allows us to re-apply components of these experiences to new situations that share some, but not all, elements. For example, we might reason that cooking a different dish at a friend's house would similarly begin by raising the heat and adding ingredients while chatting, or predict that even when driving a car, after the weather report on the radio, traffic news will follow. This reasoning is only possible through the second, more flexible, way of learning.

Despite its ubiquity, we know little about the mechanisms that enable humans to adapt to their on-going experience, where this comprises environmental dynamics generated from multiple simultaneous subprocesses (but see Conway & Christiansen, 2006 investigating behavior in response to mixed, but temporally well-separated stimulus sequences produced by two artificial grammars). In previous work, we found evidence for generalization of subprocesses that involved parsing complex tasks into component "building blocks". Based on computational modeling and behavioral evidence, we proposed a cognitive-computational mechanism that structural mapping of previous experience onto new experiences facilitates generalization of knowledge to solve new tasks (Luettgau et al., 2023). While these data extended on previous behavioral findings regarding compositionality (Lake et al., 2015, 2017; Rubino et al., 2023), we still lack a mechanistic account as to how this might be implemented at a neural level.

We conjectured that compositional generalization of structural knowledge could be enabled by a previously described mechanism that proposed abstraction of structural representations from sensory experiences within the hippocampal-entorhinal system for efficient generalization to novel situations (Behrens et al., 2018; Whittington et al., 2020, 2022). In other words, knowledge generalization involves mapping of inferred generative causes of an ongoing experience onto a structural scaffolding that supports abstractions over concrete events (Mark et al., 2020; Pesnot Lerousseau & Summerfield, 2024; Whittington et al., 2020). We propose that humans identify, neurally represent, and abstract the constitutive relational units of experience ("building blocks"). Taking the previous example, an abstracted "building block" of prior experience might entail knowledge about the sequential dynamics of cooking: apply heat → add ingredients → reduce heat → stir periodically. Such abstracted "building blocks" can then be reused as a structural scaffolding for inferring the relational units underlying new experiences. In turn, such abstract representation could facilitate a mapping of new sensory information to pre-existing structural templates, allowing flexible recombination of prior structural knowledge within new contexts. Here, we set out to provide first empirical evidence that such a mechanism supports compositional generalization of structure in humans.

A structural scaffolding account of compositional generalization entails a prediction that, across different contexts, elements of analogous subprocesses should exhibit representational alignment, indicative of a shared abstract representation that enables transfer and reuse of experiential “building blocks”. To test this, we developed a sequence learning paradigm based on the graph theoretical concept of graph factorization, emulating environments whose dynamics are the product of multiple simultaneous subprocesses. Utilizing product graphs (Cartesian product), where the joint dynamics of two or more graph factors create complex product state space dynamics that can be factorized (and simplified) into constitutive factors, provides a highly productive test bed for probing neural representation of multiple (independent) simultaneous evolving subprocesses. Examples of such graph products and their factorizations, from a vast space of possible combinations, are provided in Figure 1A. Here we invoke a specific understanding of the term *compositional*, inspired by prior usage in cognitive-computational neuroscience (e.g., Barron et al., 2013; Rubino et al., 2023; Schwartenbeck et al., 2023; alternative notions of *compositional* have been suggested in the literature, e.g., recursive operations over hierarchies (O’Donnell et al., 2009, 2011), hierarchical function composition, or rich forms of semantic composition in language to create complex concepts (Murphy, 1988)).

Using magnetoencephalography (MEG) we found that distinct sensory stimuli showed greater neural similarity when they adhered to the same underlying abstract relational dynamics, while stimuli in analogous “dynamical roles” could be successfully decoded across entirely different task phases. Decoding strength for these dynamical roles related to transfer learning success selectively for structural components that were experienced consistently across learning contexts. On this basis we suggest an emergence of abstract knowledge representations from prior experience enables reuse of distinct structural components for inference and generalization in novel contexts.

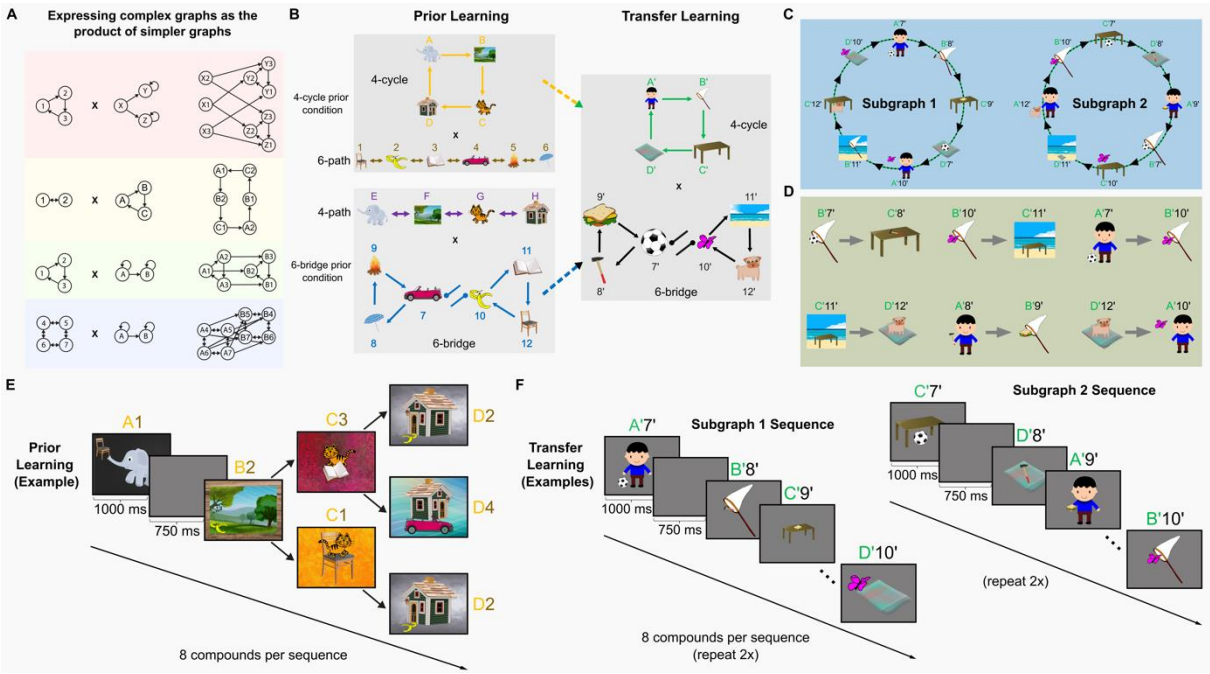


Figure 1. Experimental design

A) Exemplary product graphs (Cartesian product, right) selected from the vast array of possible graph products that can be constructed from two simpler graph components (left). The graph product of two or more simple graph components results in complex product state spaces, which can be decomposed (and simplified) into their constituting components.

B) Graph factors used in prior learning (left) and transfer learning (right)

C) During the sequence learning task (prior and transfer learning), graph components were combined into compound graphs (product graphs) by combining their images. For illustration, we only show the parts of the compound graph for transfer learning that participants experienced during sequence presentation, created from the 4-cycle and 6-bridge graphs shown in panel B (right). Note that the 6-bridge graph did not feature a truly bi-directional edge between the central nodes (7//7' and 10/10'), but rather a "refractory" edge, indicating that it was not possible to transition back and forth within one step between central nodes of the graph. In both phases of the task (prior and transfer learning), the compound graphs consisted of 24 compound images. During sequence presentations in transfer learning, participants only observed a subset of transitions from the compound graph (e.g., the 16 depicted transitions, forming two disjoint subgraphs). The remaining possible compound graph transitions formed a held-out set (see examples in panel D). This allowed us to test participants' understanding of the subprocesses generating the observed sequences. The same principles applied to prior learning as well.

D) Examples of held-out transitions used in transfer learning between compound images. Note that these transitions entailed some compounds which were observed (as part of other transitions) during sequence presentations.

E) Example sequence generated by the product graph used during prior learning. In both prior learning and transfer learning, participants were presented with sequences of meaningful compound holistic images, such as a cat on a chair or a house with a car. After each sequence presentation (consisting of 8 compound stimuli), participants' understanding of the temporal dynamics was assessed in two ways, as detailed in Figure 2. The example sequence involved a simultaneous traversal in the yellow graph factor and in the gold graph factor, as shown in panel B (left, top).

F) Example sequences produced by the subgraphs of the product graph used in transfer learning involving a simultaneous traversal in the green graph factor and in the black graph factor, as shown in panel B (right). Subgraph sequence 1 (left) was presented two times, after which subgraph sequence 2 (right) was presented twice.

Results

Participants (*N* = 51) viewed sequences of compound images comprising two elements, each image depicting a meaningful holistic scene (task design details in Figure 1 and 2). During a prior learning phase, using a between-subjects design (Fig. 1E), participants in a 4-cycle condition experienced sequences generated from the product of a 4-state cycle graph factor and a 6-state path-graph factor (Fig. 1B top panel). Likewise, participants in a 6-bridge prior condition saw compound image sequences produced by combining a 4-state path-graph factor and a 6-state bridge graph factor (Fig. 1B bottom panel). The incidental spatial relationship between the two features of each compound stimulus was kept constant per participant throughout the experiment and was not task relevant. The task design enabled us to first deduce behavioral evidence that subjects extract and reuse subprocess dynamics, in this case the underlying graph structures, replicating and extending previous findings from our research group (Luettgau et al., 2023).

The prior learning phase was followed by a transfer learning phase, identical for both conditions (Fig. 1F), but now compound image sequences were composed from entirely novel elements. The latter were the product of a 4-state cyclic graph factor and a 6-state bridge graph factor, respectively. The between group design meant that, for each participant, only one of the two experienced graph factors was common across both prior and transfer learning (e.g., in the 4-state cycle prior condition, the 4-state cycle was present in both prior and transfer learning, Fig. 1B).

In both task phases (i.e., prior and transfer learning), participants experienced sequences of a subset of all possible transitions between compound images (Fig. 1C and D illustrate this as an example for the transfer learning phase product graph). During transfer learning participants observed 16 transitions between compound images from two disjoint subgraphs of the compound graph (Fig. 1C, 8 unique compound images per subgraph). The presentation of these transitions alternated between two presentations of subgraph sequence 1 (Fig. 1F, left), and two presentations of subgraph sequence 2 (Fig. 1F, right). To avoid learning “impossible” transitions between the terminal element of one sequence to the first element of the other we explicitly instructed participants regarding the start of a new sequence presentation. The remaining eligible transitions between compound images (based on transitions on the underlying graph components) formed a held-out set of transitions between compound images (see examples in Fig. 1D). Leveraging held-out transitions enabled us to test knowledge for unobserved transitions between compound stimuli, at each task phase (i.e., prior and transfer learning). In essence this provided an assay of whether there was reuse of a graph

factorization (i.e., knowledge of the individual graph factors producing the experienced compound sequence (Fig. 1E/F)).

During prior and transfer learning, "experience probes" tasked participants to predict upcoming states that followed one of the compound images from the experienced subgraph (Fig. 2C, left column). Likewise, in both prior and transfer learning "inference probes" (Fig. 2C, right column) tested for an ability to make accurate predictions about held out set transitions. In both experience and inference probes, participants were tasked to choose between the correct next compound image and a lure compound image. The lure compound image always matched the correct compound image with respect to one of the component individual images (e.g., D'12' -> A'10' or B'10'), allowing us to test knowledge pertaining to each of the graph factors separately (here, the knowledge that in the 4-state graph factor component D' is followed by A' rather than B'). Probe questions were carefully designed in such a way that did not allow to infer the structure from these questions (see Methods for details).

Abstracting and reusing subprocesses

Our central behavioral prediction was that individuals would abstract experienced subprocesses (i.e., graph factors) from prior learning and reuse this knowledge during transfer learning, benefiting from recurrence of a previously experienced subprocess. Thus, at transfer learning the 4-cycle prior condition subjects should perform better on 4-cycle probes compared to the 6-bridge prior condition and vice versa for the 6-bridge prior condition. This predicts a disordinal interaction effect between probed size and prior learning condition on accuracy. Thus, during transfer learning we tested whether the probability of being correct varies as a function of prior learning condition (4-cycle prior condition vs 6-bridge prior condition) and probed size (probes testing knowledge of 4-state graph factor vs 6-state graph factor) (Fig. 2D).

To model candidate processes that might generate the observed transfer learning behavioral data, we defined four Bayesian multilevel GLMs (GLM1 - GLM4, Eq. 1-4). The most parsimonious model (GLM4, see Methods for model comparison) indicated the effects for probed size and prior learning condition on accuracy were qualified by a three-way interaction effect between these two variables and probe type (experience vs inference). We therefore conducted post-hoc GLMs featuring main effects for probed size, and prior learning condition and an interaction effect between these two variables separately for both probe types. Within these post-hoc models, we observed non-zero interaction effects between probed size and condition (experience probes: $\mu_{SIZExCOND} = .79$, credible interval (CI) = [.42; 1.17], Fig. 2E top row, third panel; inference probes: $\mu_{SIZExCOND} = .38$, CI = [.02; .75], Fig. 2E bottom row, third panel).

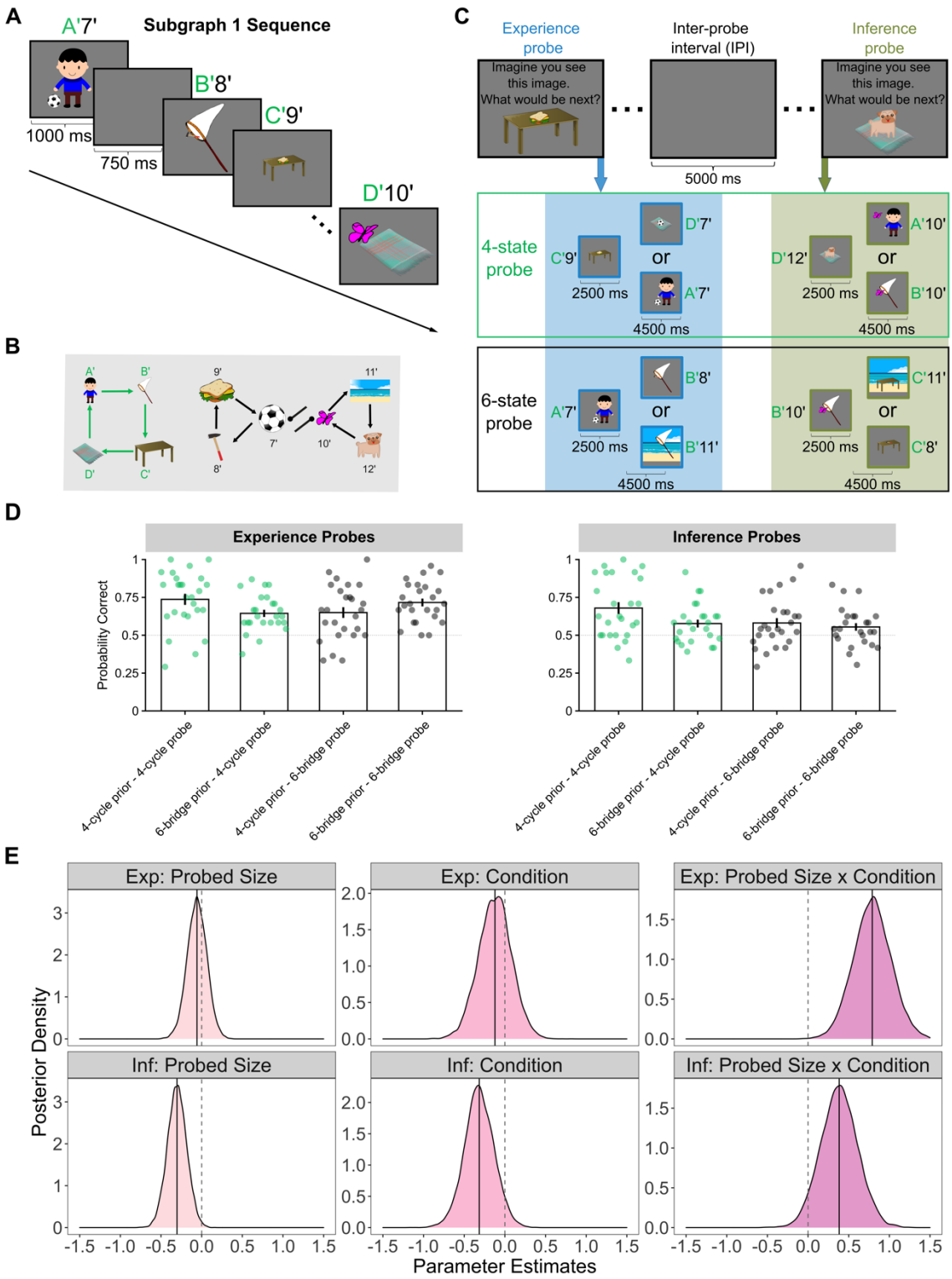


Figure 2. Behavioral task and transfer learning performance.

A) Example sequences produced by subgraph 1 of the product graph used in transfer learning.

B) The compound image sequence in panel A was generated from a simultaneous traversal of two *graph factors,* by combining their images (reproducing Fig. 1F for convenience). The example sequence was generated from a 4-state cyclic graph factor and a 6-state bridge graph factor (reproducing Fig. 1B for convenience).

C) Exemplar schematic of experience and inference probes during transfer learning, respectively. During both prior and transfer learning, following presentation of an entire compound image sequence, participants were probed to predict upcoming states that adhered to one of the 16 experienced compound image transitions (experience

probes, left column). Additionally, an inference probe (right column) tested participants' ability to infer and accurately predict held-out transitions (Fig. 1C). In both probe questions, participants saw a r compound image and were asked: "Imagine you see this image. What would be the next image?". Two compound images – correct next compound image (eligible based on the graph structure) and a lure compound image (non-eligible) – were presented as choice options. The lure compound image always matched the correct option on one of the two individual images composing the compound image (e.g., D'12' -> A'10' or B'10'). This allowed us to test knowledge of both graph factors separately (in the current example, knowledge that D' allows a transition to A' but not B') (top row: 4-state factor; bottom row: 6-state factor) underlying the observed sequence (and the held-out transitions) separately.

D) Aggregated probabilities of correct answers during the transfer learning phase, separately for experience (left) and inference probes (right) as a function of probed size (4-state cycle (green) vs 6-state bridge (black)) and condition (4-cycle prior vs 6-bridge prior). Each dot represents the arithmetic mean of experience and inference probe performance for one participant, bars represent the arithmetic mean of the distribution, error bars depict the standard error of the mean and the dashed gray line represents chance level point estimate (probability correct = 0.5).

E) Posterior density plots for each parameter estimate (posterior mean = black line) for the best-fitting GLM. The dashed vertical line represents a zero effect. In probed size (4-cycle or 6-bridge) and probe type (experience vs inference) effects, parameter estimates below zero indicate higher accuracy in the 4-cycle probes (vs. 6-bridge probes) and experience probes (vs. inference probes), respectively. In the condition effect, parameter estimates below zero indicate higher accuracy in the 4-cycle prior condition (vs. 6-bridge prior condition). The positive interaction effect indicates selectively higher accuracy in 4-cycle probes in the 4-cycle prior condition, and higher accuracy in 6-bridge probes in the 6-cycle prior condition.

Disentangling these effects using post-hoc contrasts [4-cycle prior condition – 6-bridge prior condition], showed that experience probes performance was higher for 4-cycle probes in the 4-cycle prior condition (vs 6-bridge prior condition; Fig. 2D left panel, green dots; mean difference of 10,000 posterior samples: .36, CI = [.17; .56]) whereas for 6-bridge probes, performance was superior in the 6-bridge prior condition (vs 4-cycle prior condition; Fig. 2D left panel, black dots; mean difference of 10,000 posterior samples: -.23, CI = [-.42; -.04]). In inference probes, performance on 4-cycle probes was higher in the 4-cycle prior condition (vs 6-bridge prior condition; Fig. 2D right panel, green dots; mean difference of 10,000 posterior samples: .38, CI = [.20; .56]). For 6-bridge probes, there was no evidence for a performance difference between conditions (4-cycle prior – 6-bridge prior; Fig. 2D right panel, black dots; mean difference of 10,000 posterior samples: -.11, CI = [-.26; .04]). Posterior distributions for the other parameter estimates of these GLMs are depicted in Figure 2E. Splitting behavioral results by performance levels – high performers (>50th percentile of overall prior and transfer learning performance distribution) vs low performers (<=50th percentile) – yielded similar results (Supplementary Fig. 1).

These behavioral findings replicate, and extend, previous results from a larger cohort using a similar design (Luettgau et al., 2023). The 6-bridge experience probe behavior utilized here indicates subjects not only extract and generalize subprocess dynamics for cyclical structures, but also for more complex dynamics (bridge graphs). The observed pattern is not accounted for by general practice or non-specific meta-learning related improvement effects (as such accounts would predict similar performance for both probed sizes across conditions). Instead,

the results converge with our previous findings (Luettgau et al., 2023) showing that participants extract experiential subprocess dynamics and, where there is a generative consistency across both task phases, reuse these during transfer learning.

Structural mapping of subprocesses for compositional generalization

Behavioral evidence that participants transfer knowledge across experiential contexts led us to next test a specific prediction that neural representations of experiential "building blocks" provide a structural scaffolding for processing new incoming information. This structural scaffolding account predicts elements of analogous subprocesses across different experiential contexts should show representational alignment, reflective of a shared abstract representation. To test this, we recorded MEG signals during localizer tasks from before and after sequence learning (PRE and POST localizer task, Fig. 3A). We reasoned that the predicted representational alignment of experiential "building blocks" should be reflected in a task-induced (PRE to POST) increase in neural similarity between distinct stimuli as a function of sharing the same (compared to different) underlying relational structure, in this case adhering to identical subprocess dynamics (Fig. 3B). To test this, we performed a representational similarity analysis (RSA) (Kriegeskorte et al., 2008) on MEG data (N = 50, one subject from the behavioral sample was excluded due to MEG artifacts) from PRE and POST localizer task sessions (Fig. 3A).

As a preliminary step, to denoise the sensor-level neural signal, we used a dimensionality reduction approach (principal components analysis, PCA, see Methods for details), estimating the unique PC activity related to each stimulus by regressing the principal components (20 PCs per subject) of neural data onto a localizer task-specific design matrix (stimulus identity). Using pairwise absolute differences as a distance metric, we computed the neural dissimilarity of all stimulus-specific PC activity patterns (for each PC separately), generating 20 (one for each PC) symmetrical [20 features × 20 features] Representational Dissimilarity Matrices (RDM) at each time point (-200 to 800 ms relative to stimulus onset) to generate an estimate of task-induced PRE-POST changes (POST minus PRE) in representational similarity. We then computed Kendall's tau between theoretical RDMs and PRE-POST differences in neural RDMs to quantify a representational similarity change at each time point, where this relates to a putative abstract representation of subprocesses underlying experiences (separate design matrices for the 4-state graph factor and 6-state graph factor, Fig. 3C). We focused our analysis on the top 3 PCs (PC1-3), as these explained the greatest amount of unique variance in neural data.

Following transfer learning, we observed a qualitative pattern within PC2 indicative of increased neural similarity between stimuli that share the same underlying subprocess across task phases (see PC1 and PC3 in Supplementary Fig. 2B). A group-level GLM (including condition, graph factor main effects and a condition x graph factor effect) revealed that highly distinct stimuli adhering to the same abstract relational structure (4-state cycle and 6-state bridge graphs, respectively), but presented unrelatedly during prior or transfer learning phase, manifest a significantly greater neural similarity (Fig. 3D). For illustration purposes, we computed the average difference between representational similarity metrics for both prior learning conditions (4-cycle prior condition – 6-bridge prior condition), separately for each of the graph factors (see Fig. 3C). For PC2, there was a statistically significant interaction effect of condition x graph factor covering 500 – 600 ms post-stimulus onset (corrected for multiple comparisons using permutation tests, empirical null distribution generated by shuffling the labels of the theoretical RDM 1000 times and correcting for testing the top 3 PCs, $P_{FWE} < 0.0167$, see Methods for details).

Post-hoc tests of the condition differences for each graph factor indicated that within the time period where we observed a significant condition x graph factor effect, participants in the 4-cycle prior condition showed higher positive change in stimulus representational similarity that mapped on the 4-state cycle graph factor across both task phases, by comparison to participants in the 6-bridge prior condition (who experienced a 4-state path-graph in prior learning, Fig. 3D, green line), and vice versa for the 6-bridge condition who experienced a 6-state path-graph in prior learning (Fig. 3D, black line; dark green line in Fig. 3D represents time points at which representational similarity changes for both post-hoc contrasts are significant). Performing the former analysis without first controlling for similarity across graph factors, and across the stimulus sets (e.g., similarity between prior learning 4-state graph factor stimuli and transfer learning 6-state graph factor stimuli), yielded highly similar results. These results are not explained by participants simply learning to differentiate which of the 10 observed features pertain to the 4- or 6-state graph factors, since this predicts only a main effect of graph factor on neural similarity (features within one graph factor being represented more similarly than across graph factors), but not an interaction effect of condition x graph factor.

These findings indicate a task-induced, learning-related, increase in neural similarity for stimuli conditional on their mapping onto identical abstract subprocesses shared across task phases. This suggests a learning-related representational alignment between stimuli by virtue of adhering to a common abstract template (subprocess structure), supporting a central prediction of a structural scaffolding mechanism.

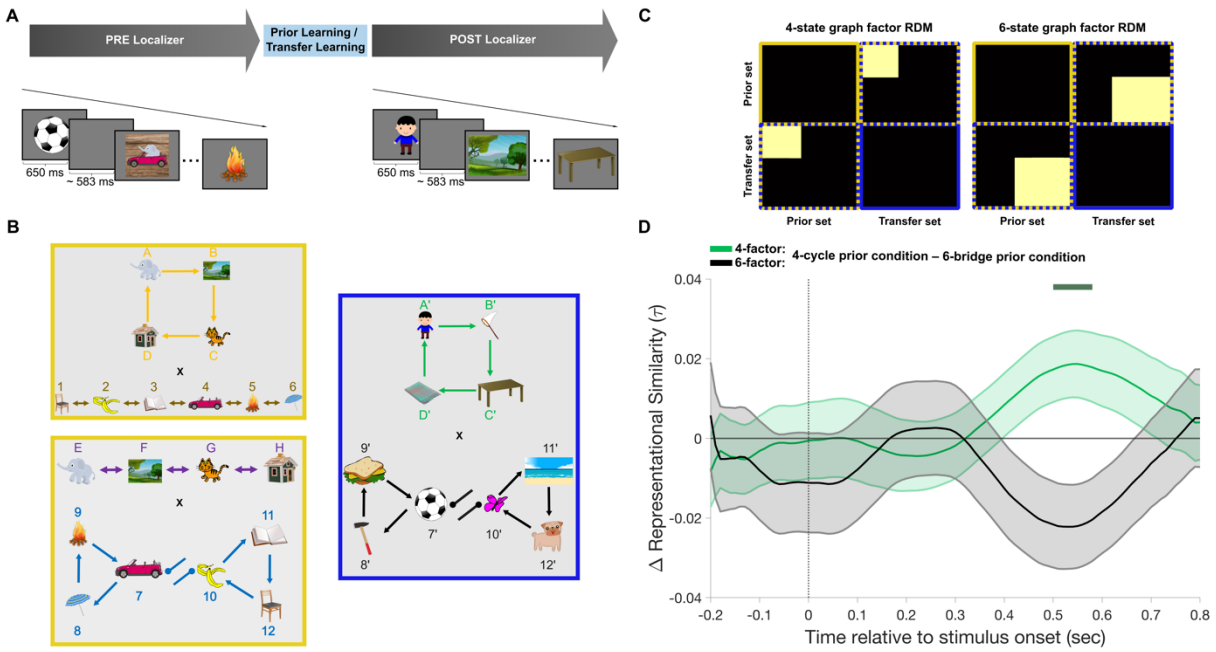


Figure 3. Neural evidence for structural mapping between subprocesses.

A) Participants performed two stimulus localizer tasks, once before and one after (PRE and POST localizer) the main experimental task (prior and transfer learning). These tasks comprised multiple presentations of all features used in prior and transfer learning (PRE and POST). A representational similarity analysis (RSA) on stimulus-evoked neural activity recorded during the two localizer tasks (PRE and POST localizer) for the two feature sets that comprise the compound images used in prior and transfer learning at time-points from -200 to 800 ms relative to stimulus-onset.

B) Feature sets in prior learning phase (yellow frame, left) and transfer learning phase (blue frame, right).

C) The presented theoretical representational dissimilarity matrices (RDM) were regressed onto the empirical neural similarity matrices – separately for the 4-state graph factor features and the 6-state graph factor features. Lighter colors denote higher values. Note that the hypothesized similarity changes are only computed across the two feature sets used in prior learning (yellow) and transfer learning (blue), but not within each set, avoiding confounding task-related changes with feature similarity unrelated to abstract subprocess structure.

D) Time courses of condition differences (4-cycle prior condition – 6-bridge prior condition) on 4-state graph factor (green) and 6-state graph factor (black) representational similarity changes in PC2 (see PC1 and PC3 in Supplementary Fig. 2B) from PRE to POST localizer. Shaded error bars indicate standard errors of difference. Dark green line represents time points at which representational changes for the condition x graph factor interaction effect and the corollary post-hoc contrasts (two-sample t-tests) exceed the maximum 98.33th percentile value of the null distribution across all timepoints (empirical null distribution generated by 1000 shuffles to the rows and columns (stimulus labels) of the theoretical RDM, see Methods for details), indicating statistically significant effects corrected for multiple comparisons at $P_{FWE} < 0.0167$ (correcting for testing the top 3 PCs, see PC1 and PC3 in Supplementary Fig. 2B). Horizontal line indicates a null effect.

Neural successor feature representations

Increased neural similarity for stimuli that map to identical abstract subprocesses supports a knowledge transfer of subprocess structure across experiential contexts. However, to enable sequence learning and to acquire accuracy on probe questions participants need to acquire predictive representations for successor states/features. On this basis we next asked whether

there are neural signatures for learning and reuse of previously acquired predictive representations. Here, MEG data from the sequence learning phase allowed us to ascertain the neural dynamics of successor feature representations, enabling us to ask how prior knowledge of subprocesses impacts learning-induced changes in the neural representation of successor features.

We first focused on characterizing neural signals that reflect knowledge representation of experienced subprocess dynamics during transfer learning. We reasoned that if neural activity encodes predictive representations, then we should be able to predict successor features from a current state's neural activity pattern. To test this, we first trained classifiers on neural activity recorded for each individually presented feature during the PRE task localizer (Fig. 3A, see Methods for details). Then, as a neural metric of successor feature representation, we tested the classifiers' generalization ability at transfer learning – generating the predicted probability for the successor features for each compound stimulus presentation, separately for both the 4-cycle and 6-bridge graph factor successor features (see Methods for details on the classification scheme). Across several time points post-stimulus onset, we found predicted probabilities for successor features pertaining to both graph factors increased above averaged pre-stimulus onset predicted probabilities (Fig. 4A, N = 49 due to convergence issues of classifiers in one participant in the 4-cycle prior condition), consistent with neural representation of successor features on both experienced subprocesses after stimulus onset.

Next, we reasoned that if previous experience of structurally identical subprocesses provides a representational scaffolding for future learning, then this predicts a learning-related evolution in the mapping between representations of shared subprocesses across contexts. This should be reflected in a gradual strengthening in the representation of successor features specific to a consistently experienced subprocess. Based on this, we predicted that neural reactivation of successor features would show an enhancement across a transition from early to late stages of transfer learning, with an advantage (i.e., stronger enhancement) specific to subprocess dynamics encountered during prior learning. Using the same classification approach detailed above, we subtracted the predicted probability of the first 10 trial successor states from the predicted probability of the final 10 trials. This revealed an interaction effect of condition and graph factor on an early to late neural signal increase, at 120 ms post stimulus onset ($F_{1,47}$ = 6.16, p = .017, mixed-effects ANOVA). For the 4-cycle successor features in the 4-cycle prior condition, compared to the 6-bridge prior condition, post-hoc testing revealed a greater change from early to late trials (t_{47} = 2.12, p = .040, two-sample t-test, Fig. 4B). At the very same time point, an early to late increase of reactivated 6-bridge successor features was also significant

for the 6-bridge prior condition compared to the 4-cycle prior condition (t_{47} = 2.08, p = .043, two-sample t-test, Fig. 4B).

These findings are consistent with the notion that humans generalize prior knowledge of subprocesses in new contexts. Such compositional generalization relies on a specific, temporally evolving mapping between subprocess dynamics that are encountered consistently across experiential contexts, a mapping that affords enhanced predictive representations of successor features.

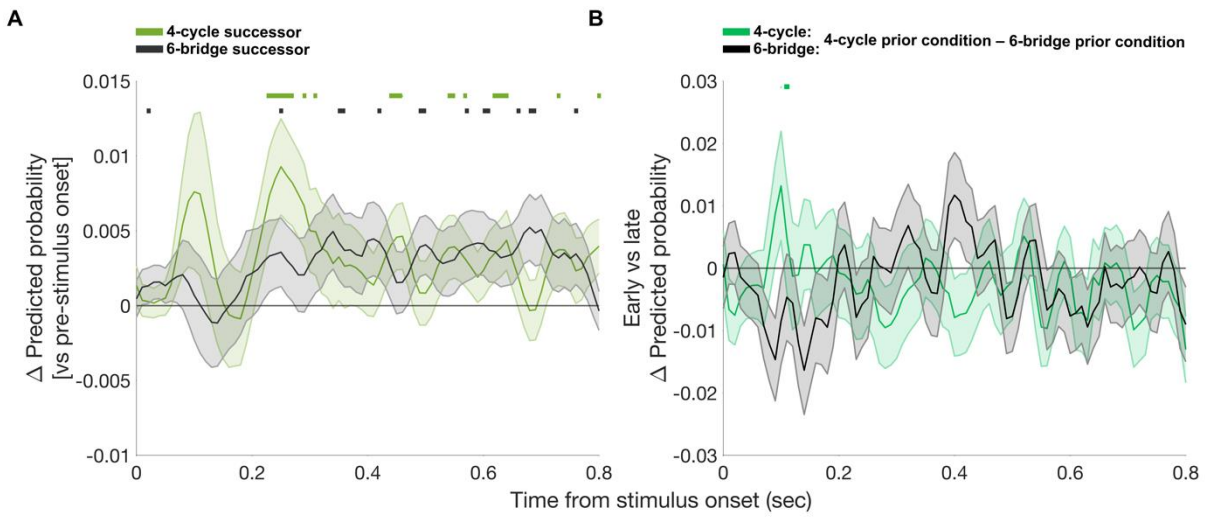


Figure 4. Temporal dynamics of successor state predictions.

A) Time courses of change in predicted probability of classifiers (reactivation) relative to average pre-stimulus onset baseline during transfer learning across both conditions, and all stimulus presentations, for the respective successor features for each compound stimulus presentation. We trained multivariate classifiers (L1-regularized logistic regressions, trained excluding the respective other feature in each compound) to distinguish between features on the 4-cycle and 6-bridge graph factor during the PRE localizer tasks and tested the compound-by-compound successor feature predictions on both graph factors separately. Depicted are average changes in predicted probability (subtracting average pre-stimulus onset predicted probability), shaded error bars indicate standard errors of the mean. Horizontal line indicates null effect (0 predicted probability change). Solid lines and dots on top index time points wherein the predicted probability is statistically significant above 0 (one-sample t-test).

B) Condition differences (4-cycle prior – 6-bridge prior condition) for time courses of predicted probability change from early (first 10 trials) to late (last 10 trials) stages of the transfer learning phase. Solid green line indicates time points where the condition x graph factor interaction effect on predicted probability changes was statistically significant (interaction effect in mixed-effects ANOVA). For these time points, the 4-cycle prior condition showed greater change on 4-cycle feature successor prediction and the 6-bridge prior condition showed greater change on 6-bridge feature successor prediction from early to late trials. Horizontal line indicates a null effect.

Abstract representation of dynamical roles in subprocesses

Learning-related changes in representational alignment of stimuli, that map to identical abstract subprocesses across task phases, in conjunction with the finding that humans generalize predictive representations to learn more effectively from new subprocess dynamics, led us to next test a more fine-grained prediction of a structural scaffolding account. Under this account, we not only expect that every element of a specific subprocess would uniformly become similar to every element of an identical abstract subprocess, but also an emergence of a fine-grained mapping of stimuli based on the precise "dynamical roles" they occupy within a graph structure. Specifically, we reasoned that if structure and dynamics of previous experiences are repurposed for new scenarios, then transitions between stimuli that share the same subprocess dynamical roles should exhibit enhanced neural representations. This predicts a higher decoding accuracy for transitions between entirely different stimuli if they share the same dynamical role. Note this can only be ascertained for the 6-state graph, as only this graph incorporates stimuli in categorically distinct dynamical roles.

To test this, we used a decoding approach on MEG data, recorded during prior and transfer learning phases, for compound stimulus presentations that adhered to specific abstract transition types in the 6-state graph factor (Figure 5A). Within the 6-state graph factor, we categorized transitions into three distinct types, reflecting their locations within the graph, e.g., transitions between boundary and intermediate graph nodes (1), transitions between intermediate and central graph nodes (2), or transitions between central and central graph nodes (3) (see Methods for details). During the prior learning phase, we trained classifiers on neural activity from stimuli that followed each of these three transition types. In the transfer learning phase, we then tested the classifier generalization to identical transition types, but where this now involves entirely different stimuli. We used two-sample t-tests to ascertain classification accuracy between participant groups who had experienced different graphs during prior learning (4-cycle prior condition experiencing a 6-path graph and 6-bridge prior condition experiencing a 6-bridge graph), with multiple comparisons correction using permutation tests (empirical null distributions generated by randomly shuffling the transition labels and computing classification accuracy 1000 times) and where significant time points were identified on the basis of classification accuracy exceeding the 95th percentile threshold ($p < 0.05$).

Following onset of a post-transition compound stimulus presentation, classification accuracy changes were evident between participants who had experienced different graphs during prior learning (6-path graph and 6-bridge graph), spanning an interval between 135 and 515 ms post stimulus-onset (Figure 5B). Thus, there was enhanced neural decodability for stimuli that

shared transition types in the 6-bridge prior condition, an effect absent in participants who had experienced the 6-path graph (4-cycle prior condition). This indicates emergence of neural similarity for stimuli whose dynamical roles adhere to the same structural dynamics across experiential contexts, supporting a more fine-grained prediction of a structural scaffolding account. While our data does not provide an unequivocal interpretation of these dynamical roles, one possibility is that individuals differentiate transitions along the 6-bridge component based on how they relate to the unique central "bridge" formed between the 2 central nodes (e.g., the transitions between nodes 7 and 10 in Fig 3B). For example, transitions can be characterized as movement outside bridge nodes (1), arriving to/moving away from bridge nodes (2) and crossing the bridge (3). This representation of dynamical roles might confer benefits to task performance – a possibility we tested next.

Dynamical role generalization and behavioral performance

Having identified a neural signature reflecting encoding of subprocess dynamics, we next asked whether dynamical role classification accuracies from the previous analysis (Fig. 5B) relate to behavioral performance during the transfer learning phase. Averaging classification accuracy over time points showing significant condition differences (Fig. 5B), we found a selective positive correlation between average classification accuracy and behavioral performance in experience probes for the 6-bridge prior condition (6-bridge prior condition: *rho* = .41, p = .041, Spearman correlation; 4-cycle prior condition: *rho* = -.16, p = .440; correlation difference: Z = 1.99, p = .046, Fisher r to z; Fig. 5C), but not for inference probes (6-bridge prior condition: *rho* < .01, p = .996, Spearman correlation; 4-cycle prior condition: *rho* = .09, p = .657; correlation difference: Z = .32, p = .752, Fisher r to z; Fig. 5D). These findings are consistent with a notion that representation of dynamical roles served by different task states has behavioral benefits. Speculatively, this knowledge representation might help individuals constrain their predictions for successive states. For example, upon transitioning to a bridge node one can predict that the successor state will involve "the other side of the bridge".

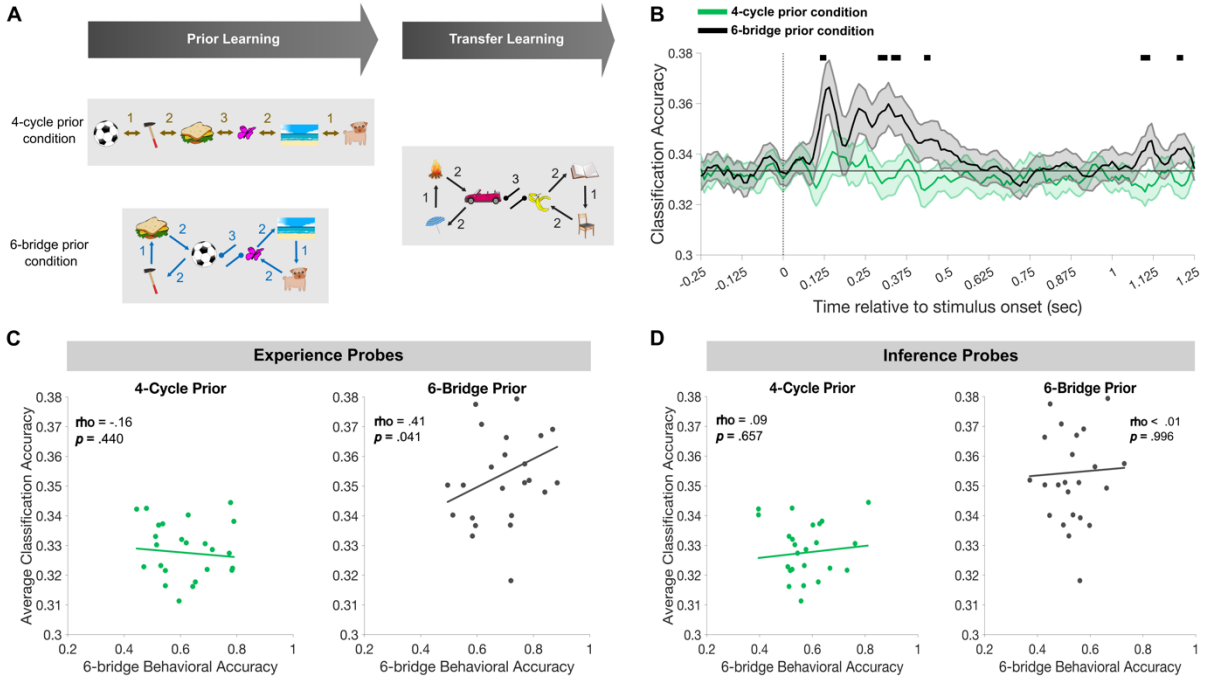


Figure 5. Abstract representation of subprocess dynamics.

A) Knowledge reuse of decomposed subprocesses, discovered during prior learning, predicts neural similarities when subprocess dynamics are shared across prior and transfer learning. During prior learning, we trained multivariate classifiers (L1-regularized logistic regressions) to distinguish between different transition types in the 6-state graph factor for 4-cycle prior condition (6-path graph, top) and 6-bridge prior condition (6-bridge graph, bottom). We then tested the classifiers' generalization ability on entirely new stimuli during transfer learning. Each number denotes a different type of transition, reflecting edges connecting graph nodes, namely boundary (1), intermediate (2), and central (3). During prior learning, classifiers were trained at post-transition stimulus presentations (e.g., in the 6-bridge prior, the classifier for transition type 2 was trained on neural activity recorded during presentations of the ball if preceded by the sandwich; or on the hammer if preceded by the ball). At transfer learning we tested the classifier's generalization performance for identical transitions, but where this now involved entirely different stimulus sets. Since training and testing were performed on independent datasets this obviates cross-validation. Note that a comparison of classification accuracy across conditions was only possible for the 6-state graph factor, as the 4-state cycle does not provide for a distinction between transition types.

B) Transfer learning time courses of 6-state graph factor transitions classification accuracy, averaged per condition based on prior experience of either 4-cycle (green) or 6-bridge condition (black). Shaded error bars indicate standard errors of the mean. Black dots represent time points over which classification accuracy differences (t-values from two-sample t-test) exceed the maximum 95th percentile value of the null distribution across all timepoints (empirical null distribution generated by 1000 random permutations of training category labels, see Methods for details), indicating statistically significant differences between the two conditions (6-bridge prior condition > 4-cycle prior condition). Horizontal lines indicate chance level decoding accuracy (.33).

C) Spearman correlations between average behavioral accuracy in 6-bridge experience probes and per participant averaged classification accuracy across all timepoints post-image onset where there is a significant condition difference between 4-cycle prior condition and 6-bridge prior condition on classification accuracy (see panel B). Each datapoint represents an individual participant. The 6-bridge prior condition (*rho* = .41, *p* = .041) showed a stronger positive correlation than the 4-cycle prior condition (*rho* = -.16, *p* = .440, correlation difference: *Z* = 1.99, *p* = .046, Fisher *r* to *z*).

D) Spearman correlations between average behavioral accuracy in 6-bridge inference probes and per participant averaged classification accuracy across all timepoints post-image onset where there is a significant condition difference between 4-cycle prior condition and 6-bridge prior condition on classification accuracy (see panel B). Each datapoint represents an individual participant. There was no evidence for a correlation difference between the 6-bridge prior condition (*rho* < .01, *p* = .996) and the 4-cycle prior condition (*rho* = .09, *p* = .657, correlation difference: *Z* = .32, *p* = .752, Fisher *r* to *z*).

Discussion

The philosophical truism of Heraclitus, that you can never step into the same river twice, reflects our everyday experience of the world. Yet even our most complex experiences share degrees of structural similarity with our past, where this often derives from a recurrence of simpler subprocesses. These recurrent world properties may explain how humans, by exploiting regularities derived from past experience as analogies, effortlessly master demands of complex and ever-changing environments (Gentner, 1983). More specifically, it is proposed that agents reuse abstract components ("building blocks") of past experience as a "scaffolding" to facilitate integration of new information (Al Roumi et al., 2021; Luettgau et al., 2023; Pesnot Lerousseau & Summerfield, 2024). Indeed, there is growing evidence consistent with a compositional (re-)use of acquired knowledge as a cognitive mechanism to explain how humans generalize knowledge from prior experiences (Barron et al., 2013; Lake et al., 2015, 2017; Luettgau et al., 2023).

Here, in the context of behavioral evidence for a parsing of subprocesses generating the dynamics of experience, that replicates previous findings from our group (Luettgau et al., 2023), we provide neural evidence consistent with an abstract representational scaffolding. The question of how such a scaffolding might be realized has been addressed in computational accounts of hippocampal and entorhinal cortex function, where grid cell activity is proposed to provide a basis set representation of abstract structure inferred from sensory experiences (Behrens et al., 2018; Mark et al., 2020; Whittington et al., 2020, 2022). We hypothesized that compositional generalization in our task might be achieved by extracting, and neurally representing, abstracted relational units of prior experience ("building blocks") that facilitates a flexible recombination in new contexts. In principle, this could be facilitated by mapping new incoming sensory information to pre-existing structural templates retrieved from memory. In testing these ideas, we provide the first human evidence that compositional generalization of structure is supported by a neural mechanism based on the formation of abstracted experiential "building blocks".

We conducted three empirical tests of the structural scaffolding neural mechanism. Firstly, we show task-induced representational alignment of unrelated stimuli, presented across different task phases and embedded in novel complex dynamics, when these adhere to identical subprocesses. Specifically, we observed a condition-specific, learning-induced, increase in neural similarity of stimuli mapping to the same abstract structural subprocesses, a finding consistent with the notion of task structure generalization (Baram et al., 2021) and neural state space alignment (Sheahan et al., 2021). Additionally, the temporal signature of the similarity change signal (peaking around 500 ms post-stimulus onset), echoes previous findings showing

emergence of an abstract structural representation, suggested to account for phenomena such as aberrant inference and delusions in disease states (Nour et al., 2021). Secondly, we found enhanced neural reactivation of successor features from early to late stages of transfer learning, preferential for subprocess dynamics participants had encountered during prior learning. This aligns with a facilitation of learning previously encountered relational patterns, even when these are now embedded by entirely different sensory stimuli, with echoes of sequence generalization in rat frontal cortex (Zhou et al., 2021). Finally, when testing a fine-grained prediction of the structural scaffolding account, we found that stimuli sharing the same dynamical role across different learned structures come to exhibit a neural similarity. Thus, neural classifiers trained during an initial task phase generalize to dynamical roles that involve entirely different stimuli if these adhere to the same abstract dynamical role within a subprocess.

The generalizability of these classifiers was positively related to experience probe behavioral performance. We acknowledge that such brain-behavioral correlation was not present for inference probes. Similarly, the group difference for the 6-bridge inference probe performance was of negligible magnitude and did not align with the robust effect observed for experience probes, where the latter result replicates our previous finding using a 6-state cycle graph, which was predicted by computational simulations using successor feature models (Luettgau et al., 2023). We speculate that the observed disparity between probe types could be due to experience probes and inference probes engaging dissociable cognitive processes – the former relying on predictive representations acquired from experience and the latter relying on an ability to perform mental manipulations to the learned representation of independent feature transitions. This may increase the likelihood of errors due to complexities in the inferred state spaces. However, the observed differences could also relate to previously reported effects of explicit knowledge on performance levels (Luettgau et al., 2023) or be driven by capacity limitations in working memory taxing larger subprocesses more strongly (Collins et al., 2017; Collins & Frank, 2012). Taken together our findings suggest striking learning and generalization abilities, whereby individuals factorize complex experiences into underlying structural elements, classify these into distinct subprocesses that rely on knowledge from past experiences and form a specialized representation of the dynamical roles these features play within subprocesses. These representations, in turn, support prediction of successive features that leads to more accurate performance.

Our neural findings suggest that building and maintenance of abstract representations of subprocesses allows factoring out the idiosyncrasies of the complex dynamics of experience. While the MEG recordings reflect an aggregate neural signal, prior literature suggests a

plausible cellular basis rests in the activity of entorhinal cortex grid cells, where these are proposed to compute invariant representations of relational structure underlying experiences, abstracted away from the sensory specifics (Behrens et al., 2018; Kazanina & Poeppel, 2023; Mark et al., 2020; Whittington et al., 2020, 2022). This representational form significantly reduces a learning computational burden by lessening a requirement for inferring the structural components determining experiential dynamics, allowing an attentional focus on the idiosyncrasies of an experience (e.g., which stimulus features are currently behaviorally relevant).

The proposed structural scaffolding mechanism also draws on ideas that include abstract neural geometries that facilitate few-shot concept learning by organizing information within structured representations (Sorscher et al., 2022). The neural signatures we identify here suggest an emergence of low dimensional representations of abstract and disentangled task variables, leveraged for generalization to new contexts, as previously documented in the hippocampal-entorhinal system (Baram et al., 2021; Bernardi et al., 2020; Courellis et al., 2024; Nieh et al., 2021), frontal cortex (Bernardi et al., 2020; Flesch et al., 2022, 2023; Zhou et al., 2021) or posterior parietal cortex (Flesch et al., 2022, 2023). Furthermore, other evidence indicates an encoding of abstract task structure (progress to goal), within rat medial frontal cortex, facilitates learning of complex behavioral sequences (El-Gaby et al., 2023). Relatedly, computational modeling indicates such representations emerge in artificial neural networks trained to perform multiple tasks (Johnston & Fusi, 2023; Yang et al., 2019). One implication of our findings is that a representational scaffolding helps to align neural state spaces that encode structure across experiential contexts. Within this account, similar state transition dynamics and dynamical roles of stimuli encountered in highly distinct experiential contexts are represented closer in a neural activity space than dissimilar ones (Bernardi et al., 2020; Courellis et al., 2024; Sheahan et al., 2021), putatively constraining the hypothesis space over abstract structural components extracted from new experiences

An important avenue for future research relates to the question of what neural mechanism supports an algorithmic implementation of the proposed structural scaffolding and how subprocess dynamics are disentangled. One plausible mechanism may involve a separation of subprocess representations by preferred reactivation at different positions in the theta cycle or by preferred reactivation coupled to distinct peaks or troughs in theta oscillations (analogously to a role of theta oscillations in vicarious trial-and-error computations, Redish, 2016). Alternatively, gradual learning, and insight-dependent decomposition of the experienced compound stimulus sequence into underlying subprocesses could be enabled by human replay, extending previous accounts that propose a role for replay in compositionality (Kurth-Nelson et al., 2023; Schwartenbeck et al., 2023), we explored the latter possibility but

did not detect reliable replay effects within the constraints of our experimental design (see Supplementary Results for discussion). Relatedly, there is also the question of how, and under what conditions, the structure of previous experiences maps to current experience. In future experiments structural similarity and size of the generalized state spaces of subprocesses in prior and transfer learning phases could be systematically varied to delineate the boundaries of mapping experiential dynamics across contexts. Future neuroimaging (fMRI) studies based on representational similarity analysis (e.g., Luettgau et al., 2022) or repetition suppression (e.g., Barron et al., 2016; Garvert et al., 2017, 2023; Luettgau et al., 2020) could contribute to pinpointing the precise functional anatomy by testing a predicted involvement of the hippocampal-entorhinal system (Baram et al., 2021; Bernardi et al., 2020; Courellis et al., 2024; Nieh et al., 2021).

A key contribution of our study is providing an empirical test of predictions of an abstract structural scaffolding as a mechanism supporting human compositional generalization. Here, we demonstrate that such a mechanism enables reuse and compositional generalization of abstracted relational units of prior experience ("building blocks"). In this account, similar state transition dynamics encountered in highly distinct experiential contexts are represented more closely in neural activity space than dissimilar ones, allowing to share components across distinct experiences. An avenue for societal impact of our work relates to the potential for augmentation of human learning, abstraction, and generalization abilities, based on leveraging simple structures and small-scale processes ("building blocks"), where deeper mechanistic insight into these abilities could even inform pedagogical strategies. Conversely, failures in efficient knowledge decomposition, involving acquisition and reuse of a representation of an abstract neural scaffolding might impact on general cognitive abilities, including the flexibility of cognition that is often considered one of the hallmarks of mental health (Huys et al., 2016; Story et al., 2024).

Methods

Participants and procedures

Participants were recruited from the local student community at University College London (UCL) via an online recruitment platform between October 2022 and May 2023. Only participants reporting the absence of past or present mental health conditions or neurological conditions were included.

A total of 61 participants were recruited with 51 included in the final behavioral data analysis sample (*N* = 1 discontinued the experiment due to a headache, *N* = 1 excluded due to a technical error, *N* = 1 excluded due to indicating a diagnosis of aphantasia in post-experimental debriefing, *N* = 3 due to repeatedly falling asleep and closing their eyes during the experimental task, *N* = 4 were excluded due to low motivation and indicating that they gave up learning the sequence during the experiment). One additional participant was excluded from MEG data analyses due to a metal artifact, leaving 50 healthy volunteers for the final MEG sample (median age: 23 years, *SD* = 3.69, range = 18 – 36, 9 males), *N* = 25 subjects per condition. We did not perform a formal power calculation to determine the sample size but instead opted to base the sample size on that used in similar MEG studies previously conducted in our lab. Participants received £10.00 per hour as compensation. The study was approved by the University College London Research Ethics Committee (reference number: 3090/004) and conducted in accordance with the Declaration of Helsinki.

After providing informed written consent subjects were prepared for the MEG scan. Participants received written instructions regarding the pre-task localizer session on the presentation screen in the MEG scanner room. Following these instructions, subjects were asked to explain the pre-task localizer. Subsequently, participants received instructions regarding the sequence learning task and then were randomly assigned to one of the two experimental conditions (see Experimental design and behavioral task). Upon completion of the sequence learning task, participants performed a post-learning localizer task. At the end of the experiment participants filled out a sociodemographic questionnaire and a post-experimental questionnaire that assessed strategies used as well as their general understanding of the task, followed by a debriefing regarding the aims and motivation behind the study.

Experimental design and behavioral task

PRE localizer

Stimuli representing states in the graphs comprised depictions of everyday objects, landscapes, and animals/humans, arranged to create meaningful holistic scenes (e.g., a boy and a ball, or a cat in a car). Images were selected from Pixabay (https://pixabay.com/), where two images were manually assembled to form one compound image using Inkscape (https://inkscape.org/). We used two non-overlapping and unrelated sets of ten images to compose the compound images. The assignment of sets and the positions of the two images used to create the compound images on the underlying graph structures producing the observed sequences were counterbalanced across subjects and conditions. This mitigated a potential confound in learning, as well as choice biases, due to idiosyncrasies inherent in the compound images. The PRE localizer task comprised 5 blocks, involving presentation of 960 or 1020 stimuli (depending on whether participants were allocated to the 4-cycle prior or the 6-bridge prior condition) in the center of the MEG presentation screen (stimulus presentation duration: 650 ms), with an inter-stimulus interval of 583 ($\pm$116) ms, drawn from a truncated, discretized gamma distribution (shape = 1.5, scale = 0.5), marked by a blank screen. The 20 distinct features presented during prior and transfer learning phase (10 features per task phase, respectively) with the 12 or 14 compound images (4-cycle prior or the 6-bridge prior condition, respectively) shown during the prior learning phase being presented 30 times each.

Each compound image was presented as foreground to a unique background image (textures) that was randomly assigned to each compound image per participant. This design feature was intended to amplify the dissimilarity between compounds that might share individual features. Participants were instructed to monitor the presented stimuli carefully. Attention to the presented stimuli was assessed based upon 10% pseudo-randomly selected stimulus presentations being used to probe subjects regarding the last presented image. Here, a probe question, “Previous image?” was presented alongside two English words, on the left- and right-hand side to the center of the screen, describing the correct last stimulus or an alternative incorrect stimulus (order randomized across stimulus presentations, duration: 4000 ms). The words describing compound images were selected such that the holistic nature of the scene (instead of the distinct features) was emphasized. Each stimulus (features and compounds) was probed equally often. Participants responded by selecting the left or right response button and if they did not respond within 4000 ms a time out warning message was presented for 1500 ms. Participants were not provided with feedback about the accuracy of their responses but received summary feedback about their overall accuracy during the last block.

POST localizer

The POST localizer task was identical to the PRE localizer task. However, only the 20 features were presented, but not the compound images shown during the prior learning phase. Thus, during the POST localizer participants were presented with 600 stimuli.

Sequence learning task

In a sequence learning paradigm (Luettgau et al., 2023) subjects were presented with sequences of compound images. In a between-subjects design, during prior learning one group of participants (condition 1, 4-state cycle prior) observed sequences of image compositions (12 compound images total, composed from 10 individual features) drawn from the product graph of a cyclic graph factor with 4 states (e.g. A->B->C->D->A …) and a path-graph factor with 6 states (e.g. 1<->2<->3<->4<->5<->6, 4 x 6 factorization). The other group (condition 2, 6-state bridge prior) saw compound image sequences (14 compound images total) produced by a bridge graph factor with 6 states (e.g., 7->8->9->7->10->11->12->10->7, …, 4 x 6 factorization) and a path-graph factor with 4 states (e.g., E<->F<->G<->H). Participants observed 54 image sequences during this ‘prior learning’ phase. Each compound image was presented in front of a unique background image (textures) that was randomly assigned to each compound image per participant.

Subsequently, during a transfer learning phase, we tested subjects on entirely new image compositions (16 compound images total, composed from 10 novel individual features) where these were generated from the product graph of a cyclic graph (4 states, e.g. A’->B’->C’->D’->A’…) and a bridge graph (6 states, e.g., 7’->8’->9’->7’->10’->11’->12’->10’->7’, …, 4 x 6 factorization). Participants observed 48 image sequences during this transfer learning. Note that the 6-bridge graphs used in prior and transfer learning did not feature a truly bi-directional edge between the central nodes (7//7’ and 10/10’), but rather a “refractory” edge, indicating that it was not possible to transition back and forth within one step between central nodes of the graph.

The nature of the factorization allowed us to present only a subset of 12 or 14 compound stimuli, respectively. Therefore, during both task phases, participants did not experience the full set of all 24 possible image compositions implied by the 4 x 6 graph factorization. A held-out set of transitions between compound stimuli (partially observed during sequence presentations, partially never observed during sequence presentations) provided a context for assessing knowledge of graph factorization.

Each sequence comprised 8 compound image presentations in the center of the screen (stimulus presentation duration: 1000 ms), interleaved by an inter-stimulus interval of 750 ms, marked by a blank screen.

Following each presented sequence, two probe questions asked participants to predict upcoming states. This assessed knowledge of either of transitions between the experienced compound stimuli in the sequence (experience probes), or their ability to infer and predict never experienced transitions between compound images, where the latter are implied by the 4 x 6 graph factorization (inference probes). We reasoned that participants could only attain above chance predictions about never experienced transitions if they had encoded a factorized representation of the true generative state space underlying the observed image compositions.

In each probe question, participants first saw a randomly selected compound image from the experienced sequence, or an entirely new compound image, with a probe question "Imagine you see this image. What would be the next image?" (duration: 3000 ms). Probe questions testing knowledge of never experienced transitions between compound images did not feature unique background images. Following the probed image, two compound images – the correct next image and a lure image – were presented as choice options on the left- and right-hand side next to the screen center. Crucially, on each probe, the lure image was matched with the correct option on one of the two features (e.g., A1 -> B2 or D2), a design feature that allowed us to test knowledge of both graph factors underlying the observed sequence separately, as well as assess knowledge about never experienced transitions between compounds. Each unique compound image in the sequence (and each never experienced transition) was used exclusively to assess knowledge of one of the two graph factors, but not both. The latter was intended to control for a possibility that participants might guess and remember the correct answer for each probe question by observing the same correct option repeatedly presented alongside different, changing lure images – allowing for the inference that the consistent option must be the correct one – even without knowing anything about the compound sequence, or correctly inferring transitions between novel compounds. Each transition within each graph factor was probed equally often in both prior and transfer learning phases.

During each probe question, participants had 4500 ms to select the left or right option by using the left or right button on two button boxes held in both hands. If they failed to respond within this time window, a timeout warning was presented for 1500 ms and the probe question was coded as a missing event. Participants did not receive feedback about whether they answered correctly or not on any given probe question. Between the experience and inference probe questions, there was a blank screen for 5000 ms (inter-probe interval). After the inference probe, the sequences proceeded.

The task comprised 8 blocks, with short breaks after every 14th or 12th (transfer learning phase) presented sequence and a longer break (5 min) between prior and transfer learning. During these breaks, participants received feedback about their average performance in the experience probe questions up until this trial. As well as assessing whether participants entertained a factorized representation of the graph factors generating the sequence of compound observations, our transfer learning task design allowed us to test a prediction that participants form abstractions of the decomposed or factorized state space, and reuse components of the abstracted generative processes (graph factors) they had encountered in prior learning in an entirely new context, the transfer learning phase. Importantly, the experimental design also implies a second type of transfer learning, other than reusing a graph factor during the second learning phase: Participants could repurpose their knowledge about the experienced transitions (as assessed by experience probes) to infer never experienced transitions on the graph factors (as assessed by inference probes).

Behavioral analyses

Data were preprocessed and analyzed in MATLAB 2023b (The MathWorks, Inc., Natick, MA, USA) and RStudio (RStudioTeam, 2019) (version 4.1.0, RStudio Team, Boston, MA) using custom analysis scripts. We defined Bayesian multilevel generalized linear models (GLM), representing different processes that might have generated the observed choice data using the R package rethinking (McElreath, 2020a, 2020b). We specified binomial likelihood functions to model the (aggregated) number of correct responses distributed as the proportion of correct probes (binomial distribution parameter *p*) among the number of probes for which a response was recorded (number of probes). We employed sampling-based Bayesian inference to estimate the posterior distribution of linear model parameters comprising different intercept and slope parameters that linearly combine to form the binomial distribution parameter *p*. In the below models, μ denotes average/group level effects ("fixed effects"), α and γ denote individual or individually covarying effects ("random effects"). We coded the probed size effect as –.5 (for 4-cycle probes) and as .5 (for 6-bridge probes). A negative value of the parameter estimate indicates higher accuracy for 4-cycle probes. The probe type effect was coded similarly (experience probes as –.5; inference probes as .5). A negative value of the parameter estimate indicates higher accuracy for experience probes. The condition effect was coded as –.5 (for 4-cycle prior condition) and as .5 (for 6-bridge prior condition). A negative value of the parameter estimate indicates higher accuracy for the 4-cycle prior condition.

For each of the below models, we specified weakly informative prior and hyperprior probability distributions (as indicated in the model specifications). Models were passed to RStan (Stan Development Team, 2020) using the function "ulam" (rethinking package). We drew 4 x 4000 samples from posterior probability distributions (4 x 1000 warmup samples), using No-U-Turn samplers (NUTS; a variant of Hamiltonian Monte Carlo) in RStan and four independent Markov chains. Quality and reliability of the sampling process were evaluated with the Gelman-Rubin convergence diagnostic measure ($\hat{R} \approx 1.00$) and by visually inspecting Markov chain convergence using trace- and rank-plots. For all models fitted we found $\hat{R} = 1.00$ for all parameters sampled from the posterior distribution. There were no divergent transitions between Markov chains for any of the models reported.

For model comparisons and to find evidence for the best-fitting model for the observed behavioral data, we used the Widely Applicable Information Criterion (WAIC). Parameter estimates were considered non-zero if the Bayesian credible interval (CI) around the parameter did not contain zero.

We defined a model space of four Bayesian multilevel (statistical models, GLM1 – GLM4, Eq. 1-4) generalized linear models (cf. Luettgau et al., 2022, 2023). Both experience and inference probe correct responses were analyzed in a joint model to reduce the number of tests of the same statistical hypothesis and to increase statistical power (a joint model should capture within-subject performance ability and correlations across both probe types more adequately and explicitly than two separate models for both probe types).

To test our main hypothesis, that participants factorized and abstracted structure underlying their experience during prior learning and reused an abstracted graph factor that was consistent across prior and transfer learning, we focused on transfer learning phase behavior. We expected an advantage in predicting 4-cycle graph factor transitions in the 4-cycle prior condition (vs 6-bridge prior condition), and vice versa, enhanced performance in 6-cycle graph factor transitions in the 6-bridge prior condition (vs 4-cycle prior condition). Formally, this pattern of responses should reflect in an interaction effect between condition and probed size. We therefore predicted that the best-fitting model for the observed behavioral data (as found in model comparison) would feature a non-zero interaction effect between condition and probed size, suggesting that this effect captures important unique variance in explaining participants' choice behavior.

GLM1. Individually varying intercepts model. This model assumes that correct responses are invariant across probed sizes, probe types, and conditions.

$$\begin{gathered} Correct_n \sim Binomial(number\ of\ probes_n, p_n) \\ logit(p_n) = \mu_\alpha + z_{subject[n]} * \alpha_{sigma} \\ \mu_\alpha \sim Normal(0,1) \\ \alpha_{sigma} \sim Half - Normal(0,1) \\ z_{subject[n]} \sim Normal(0,1) \end{gathered} \qquad \textit{(Eq. 1)}$$

where *Correct* denotes the number of correct probes, distributed as the proportion of correct probes (binomial distribution parameter *p*) among the number of probes for which a response was recorded (number of probes), *n* denotes the parameter estimate for the n-th subject. Note that the model is reparametrized to allow sampling from a standard normal posterior distribution. α_{sigma} denotes the standard deviation of the reparameterized normal distribution of the varying intercepts parameter α.

GLM2. Individually varying intercepts, main effect of probe type (*PT*), covarying intercept and probe type effect

$$\begin{gathered} Correct_{j,n} \sim Binomial(number\ of\ probes_{j,n}, p_{j,n}) \\ logit(p_{j,n}) = \mu_\alpha + \mu_{PT[j]} + \alpha_{subject[n],PT[j]} \\ \begin{bmatrix} \alpha_{n,1} \\ \alpha_{n,2} \end{bmatrix} \sim MVNormal\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, S\right) \\ S = \begin{pmatrix} \sigma_{subject} & 0 \\ 0 & \sigma_{PT} \end{pmatrix} R_{PT} \begin{pmatrix} \sigma_{subject} & 0 \\ 0 & \sigma_{PT} \end{pmatrix} \\ \mu_\alpha \sim Normal(0,1) \\ \mu_{PT} \sim Normal(0,1) \\ \sigma_{subject} \sim Half - Normal(0,1) \\ \sigma_{PT} \sim Half - Normal(0,1) \\ R \sim LKJcorr(2) \end{gathered} \qquad \textit{(Eq. 1)}$$

where *MVNormal* is a multivariate normal/Gaussian distribution, the correlation matrix R_{PT} is distributed *as LKJcorr* distribution (Lewandowski-Kurowicka-Joe distribution). We used non-centered parameters for the covariance matrix S (Cholesky decomposition) to facilitate estimation using sampling-based inference. j denotes the parameter estimate for the j-th probe type (experience or inference probe). σ are the variance parameters for the subject specific intercept parameter and probe type parameter.

GLM3. Individually varying intercepts, main effects of probed size, probe type and condition, interaction effect of probed size x condition (*SIZE x COND*), covarying intercept and probed size, probe type effect

$$\begin{gathered} Correct_{i,j,k,n} \sim Binomial(number\ of\ probes_{i,j,k,n}, p_{i,j,k,n}) \\ logit(p_{i,j,k,n}) = \mu_\alpha + \mu_{SIZE[i]} + \mu_{PT[j]} + \mu_{COND[k]} + \alpha_{subject[n],SIZE[i]} + \\ \gamma_{subject[n],PT[j]} + \mu_{SIZEx\ COND} \\ \begin{bmatrix} \alpha_{n,1} \\ \alpha_{n,2} \end{bmatrix} \sim MVNormal\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, S_{SIZE}\right) \end{gathered} \qquad \textit{(Eq. 3)}$$

$$\begin{bmatrix}\gamma_{n,1}\\ \gamma_{n,2}\end{bmatrix} \sim MVNormal\left(\begin{bmatrix}0\\0\end{bmatrix}, S_{PT}\right)$$

$$S_{SIZE} = \begin{pmatrix}\sigma_{subject} & 0\\ 0 & \sigma_{SIZE}\end{pmatrix} R_{SIZE} \begin{pmatrix}\sigma_{subject} & 0\\ 0 & \sigma_{SIZE}\end{pmatrix}$$

$$S_{PT} = \begin{pmatrix}\sigma_{subject} & 0\\ 0 & \sigma_{PT}\end{pmatrix} R_{PT} \begin{pmatrix}\sigma_{subject} & 0\\ 0 & \sigma_{PT}\end{pmatrix}$$

$$\mu_{\alpha} \sim Normal(0,1)$$
$$\mu_{SIZE[i]} \sim Normal(0,1)$$
$$\mu_{PT[j]} \sim Normal(0,1)$$
$$\mu_{COND[k]} \sim Normal(0,1)$$
$$\mu_{SIZEx\,COND} \sim Normal(0,1)$$
$$\sigma_{subject} \sim Half-Normal(0,1)$$
$$\sigma_{SIZE} \sim Half-Normal(0,1)$$
$$\sigma_{PT} \sim Half-Normal(0,1)$$
$$R_{SIZE} \sim LKJcorr(2)$$
$$R_{PT} \sim LKJcorr(2)$$

where *k* denotes the parameter estimate for the k-th condition (4-cycle prior condition or 6-bridge prior condition), i denotes the parameter estimate for the i-th probed size (4-cycle probe or 6-bridge probe)

GLM4. Individually varying intercepts, main effects of probed size, probe type and condition, interaction effect of probed size x condition and probed size x probe type x condition (*SIZE x PT x COND*), covarying intercept and probed size, probe type effect

$$Correct_{i,j,k,n} \sim Binomial\left(number\ of\ probes_{i,j,k,n}, p_{i,j,k,n}\right)$$

$$logit\left(p_{i,j,k,n}\right) = \mu_{\alpha} + \mu_{SIZE[i]} + \mu_{PT[j]} + \mu_{COND[k]} + \alpha_{subject[n],SIZE[i]} + \gamma_{subject[n],PT[j]} + \mu_{SIZEx\,COND} + \mu_{SIZExPTxCOND}$$

$$\begin{bmatrix}\alpha_{n,1}\\ \alpha_{n,2}\end{bmatrix} \sim MVNormal\left(\begin{bmatrix}0\\0\end{bmatrix}, S_{SIZE}\right)$$

$$\begin{bmatrix}\gamma_{n,1}\\ \gamma_{n,2}\end{bmatrix} \sim MVNormal\left(\begin{bmatrix}0\\0\end{bmatrix}, S_{PT}\right)$$

$$S_{SIZE} = \begin{pmatrix}\sigma_{subject} & 0\\ 0 & \sigma_{SIZE}\end{pmatrix} R_{SIZE} \begin{pmatrix}\sigma_{subject} & 0\\ 0 & \sigma_{SIZE}\end{pmatrix}$$

$$S_{PT} = \begin{pmatrix}\sigma_{subject} & 0\\ 0 & \sigma_{PT}\end{pmatrix} R_{PT} \begin{pmatrix}\sigma_{subject} & 0\\ 0 & \sigma_{PT}\end{pmatrix} \qquad \textit{(Eq. 4)}$$

$$\mu_{\alpha} \sim Normal(0,1)$$
$$\mu_{SIZE[i]} \sim Normal(0,1)$$
$$\mu_{PT[j]} \sim Normal(0,1)$$
$$\mu_{COND[k]} \sim Normal(0,1)$$
$$\mu_{SIZEx\,COND} \sim Normal(0,1)$$
$$\mu_{SIZExPTxCOND} \sim Normal(0,1)$$
$$\sigma_{subject} \sim Half-Normal(0,1)$$
$$\sigma_{SIZE} \sim Half-Normal(0,1)$$
$$\sigma_{PT} \sim Half-Normal(0,1)$$
$$R_{SIZE} \sim LKJcorr(2)$$
$$R_{PT} \sim LKJcorr(2)$$

GLM comparison

A multilevel GLM featuring all of the above effects and a three-way interaction effect of probed size x probe type x condition (GLM4) showed slightly better model fits than a multilevel GLM featuring individually varying intercepts, main effects of probed size and probe type, as well as covariation with the individually estimated intercept parameters, a main effect of condition, and an interaction effect of probed size x condition (GLM3) (WAIC values ($\pm$ standard error): GLM4 = 1015.2 ($\pm$21.25), GLM3 = 1020.2 ($\pm$22.11)). All other GLMs explained the data less well than GLM4 (WAIC values ($\pm$ standard error): GLM2 = 1101.4 ($\pm$33.04), GLM1 = 1151.9 ($\pm$38.60)). Crucially, there was strong model evidence that GLMs featuring varying intercepts alone (e.g., GLM1), representing the alternative hypothesis of participants entertaining a compound representation of their experience, was much worse explaining observed behavioral data.

The model comparison suggests that the effects for probed size and prior learning condition on behavioral accuracy were qualified by a three-way interaction effect between these two variables and probe type (experience vs inference) (GLM4). We thus followed up on this analysis by using post-hoc GLM3 featuring main effects for probed size, and prior learning condition and an interaction effect between these two variables separately for both probe types (excluding the probe type main effect)

MEG acquisition and preprocessing

MEG (magnetoencephalography) was recorded at 1200 Hz using a whole-head 275-channel axial gradiometer MEG system (CTF Omega, VSM MedTech), while participants sat upright. 3 sensors were not recorded due to excessive noise. Preprocessing was conducted separately for each recording block, identical to previous studies from our laboratory (e.g., (Liu et al., 2019; Nour et al., 2021). Sensor data were high-pass filtered at 0.5 Hz to remove slow-drifts. Data were then down sampled to 100 Hz. Noisy segments and sensors were automatically detected and removed before independent component analysis (ICA). ICA (FastICA, http://research.ics.aalto.fi/ica/fastica) was used to decompose the sensor data for each session into 150 temporally independent components and associated sensor topographies. Artifact components (e.g., eye blink and heartbeat interference) were automatically classified by consideration of the spatial topography, time course, kurtosis of the time course and frequency spectrum for all components (Liu et al., 2019; Nour et al., 2021). Artifacts were rejected by subtracting them out of the data. Before cutting the MEG data into individual stimulus presentations, we re-aligned triggers to a photodiode signal that was recorded in response to each stimulus presentation, to ensure high precision of neural time series. All MEG analyses were performed on the filtered, cleaned MEG signal from each sensor at whole-brain sensor level, in units of femtotesla.

MEG data analysis

Representational similarity analysis

We leveraged representational similarity analysis (RSA) (Kriegeskorte et al., 2008) to investigate task-related changes in abstract graph factor representations from PRE to POST localizer. To this end, for each session we z-scored the pre-processed MEG data over all stimulus presentations, for each sensor and from -200 to 800 ms relative to stimulus-onset. To denoise the data and reduce dimensionality, the z-scored sensor x time x stimulus presentations data was subjected to principal components analysis (resulting in a PC [20] x time x stimulus presentations data matrix). We computed neural dissimilarity for stimulus-specific PC activity using pairwise absolute differences as a distance metric. Specifically, we regressed the neural data onto a session-specific design matrix, denoting the stimulus label of each stimulus presentation (dummy coded). We used the resulting [stimulus x 1] vector of regression weights, as an estimate of the unique PC activation related to each stimulus. Repeating this procedure over all PCs yielded a [PC x stimulus] matrix at each time point, which was spatially pre-whitened prior to calculating pairwise absolute differences as distance between the PC patterns for each pair of stimuli. This generated a symmetrical [20 × 20] Representational Dissimilarity Matrix (RDM) at each time point. We conducted this procedure in the same manner over both PRE and POST localizer tasks, enabling us to calculate the task-induced PRE-POST representational similarity change at each time point. We regressed out across set confusion (e.g., prior learning 4-state graph factor and transfer learning 6-state graph factor) to control for general effects of set similarity. Finally, on the residuals of this regression, we computed Kendall's tau to quantify the variance in representational similarity change at each time point that was uniquely explained by an abstracted representation of the graph factors (separate design matrices for the 4-state graph factor and 6-state graph factor). The representational similarity change was smoothed over time with a 25 ms Gaussian kernel prior to computing Kendall's tau.

We focused our analyses on the top 3 PCs, as these explained most variance in the data. We used nonparametric tests to identify timepoints with significant positive evidence for the hypothesized effects, correcting for multiple comparisons over time. Specifically, for a given effect (e.g., 4-state graph factor similarity) at each time point we conducted a separate one-sample t-test on the representational similarity change over participants, to obtain evidence for an effect > 0 (i.e., t-value). To correct for multiple comparisons, we then repeated this procedure for 1000 permutations, on each occasion shuffling the rows and columns (stimulus

labels) of the theoretical RDM prior to computing Kendall's tau (shuffled order consistent across time within each permutation to preserve temporal smoothness in visually evoked neural data). We then extracted the maximal t value for the group mean effect in each permutation to generate an empirical null distribution for the hypothesized effect in question. A suprathreshold statistical value in the true data was deemed significant at $P_{FWE} < 0.0167$ (correcting for testing the top 3 PCs) if it exceeded the 98.33rd percentile of the empirical null distribution. Similarly, we conducted permutation tests comparing the two groups (two-sample t-tests) and computed a group x graph factor interaction effect.

Successor feature prediction

This analysis aimed at assessing evidence for neural representation of successor features in both subprocesses.

We employed L1-regularized (LASSO) logistic regression classifiers to MEG data recorded during stimulus presentations in the PRE task localizer and transfer learning phase. Individually optimized L1 penalty hyper-parameter values (λ) per participant were obtained using leave-one-run-out cross-validation (5-fold CV, training on four of the localizer blocks and testing classifier performance on the remaining block) individually on neural activity recorded at all individually presented transfer learning phase features during the PRE task localizer. Optimized hyper-parameters were used to train classifiers per participant on the transfer learning phase features presented during the PRE task localizer. Since we employed a one vs all classification scheme and every compound stimulus presented during the transfer learning phase was composed of two features, we reasoned that the classifier representing the other (irrelevant) feature in the compound stimulus would show the highest predicted probability. We sought to control for the confounding influence of the other stimulus (of no interest) contained in the compound stimulus by training and testing classifiers excluding the respective other stimulus, resulting in n * n-1 (10 * 9 = 90) classifiers. We then applied the trained classifiers to test their generalization abilities in the transfer learning phase – generating the predicted probability for each successor feature for each subprocess separately, on each compound stimulus presentation. Since all classifiers showed some level of predicted probability, we used the raw predicted probability as a neural metric of successor state representation (instead of computing the classification accuracy for the next state only). No cross-validation was performed since training and testing were performed on independent datasets.

We analyzed neural activity in a time window from 0 – 800 ms post-stimulus onset for each presented compound stimulus. This interval was chosen since the PRE task localizer presented each stimulus for 650 ms. The 800 ms post-stimulus period was shorter than the 1000 ms image display duration in transfer learning – we therefore make no claims about successor state reactivations that might have potentially occurred between 800 and 1000 ms post-stimulus onset.

To quantify the neural representation of successor features, we subtracted the averaged pre-stimulus onset (-200 – 0 ms) predicted probabilities on a given stimulus presentation from the post-stimulus predicted probabilities – providing a neural metric for feature reactivation. We conducted one-sample t-tests against 0 (null effect) predicted probability, jointly for both prior learning conditions at each time point post-stimulus onset for the 4-cycle and 6-bridge subprocess successor feature representation.

Similarly, for the early vs late change comparison across prior learning conditions and graph factors, we computed two-sample t-tests and computed a group x graph factor interaction effect (mixed-effects ANOVA).

Decoding of dynamical roles

The following analysis aimed to assess neural evidence for an abstract representation of transitions in the 6-state graph factor. To ensure higher quality MEG data during prior and transfer learning, we only retained sensors that were not rejected in any of the experimental runs of both phases.

We investigated different transition types between images on the 6-state graph factor (i.e., the 6-state path-graph factor or the 6-state bridge graph factor) (Figure 5A). While the 6-state graph factor permits the categorization of transitions into three types, such differential labeling and classification is not possible for the 4-cycle graph factor, where all transitions are structurally identical with respect to the outbound and inbound node. In the 6-state graph factor, we grouped each transition between features into their three distinct types: boundary (1), intermediate (2), and central (3). Boundary transitions represent the transitions at the graph factor's peripheries, occurring between nodes at extreme ends. Intermediate transitions represent the transitions from boundary nodes towards the center, bridging the boundary and central nodes of the graphs. Center transitions represent transitions within the two central nodes of the graph.

We analyzed neural activity in a window from 250 ms pre- to 1250 ms post-stimulus onset for each post-transition stimulus. The 1250 ms post-stimulus period encompassed the 1000 ms image display duration. Neural activity during stimulus presentations following each transition label was used to train L1-regularized logistic regressions to classify transitions during the prior learning phase (one classifier per transition type, trained in a one-vs-all fashion). Classifiers were then tested for generalization in the transfer learning phase with different stimulus sets used in both phases. No cross-validation was performed since training and testing were on independent datasets.

In the testing phase, we applied the trained classifiers to predict each transition type given the neural activity. To compute classification accuracy, at each sample time point, we selected the classifier with the highest predicted probability and compared the maximum predicted class to the correct ground truth label class. Given that there were three types of transitions, the chance level for classification was 33%.

Further, we conducted two-sample t-tests at each time point to compare classification accuracy between participants in 4-state cycle prior and 6-state bridge prior condition. To correct for multiple comparisons, we implemented a permutation test, repeating the process 1000 times. During each permutation, transition labels were randomly reassigned to neural activity, and the decoding process was reiterated. We derived t-statistics for group comparisons at 151 time points from all permutations, identifying time points where classification accuracy significantly surpassed the significance level ($p < 0.05$). Using the permutation distribution of t-statistics at each timepoint, we calculated the 95th percentile and selected the maximum value across all timepoints as our final threshold for significance. We then identified time points where the classification accuracy significantly exceeded this threshold. Specifically, we identified significant time points ($p < 0.05$) where the two groups differed after stimulus onset.

Separately for each prior learning condition, we computed the classification accuracy averaged across the time points showing a significant condition difference. Subsequently, we computed the Spearman correlation coefficient for the condition-averaged classification accuracy at the timepoints of significant condition differences and the 6-bridge probe performance (separately for experience and inference probes) for each condition separately. To assess significant differences between correlation coefficients, they were transformed to z-scores and compared (Fisher's r to z transformation).

Code Availability

All code used to generate the results and figures in this paper will be made on GitHub upon publication.

Data Availability

Behavioral data and MEG data used to generate the results and figures in this paper will be made available on GitHub upon publication.

Acknowledgements

Financial acknowledgments: RJD (Wellcome Investigator Award, 098362/Z/12/Z). The Max Planck UCL Centre is supported by UCL and the Max Planck Society. The Wellcome Centre for Human Neuroimaging (WCHN) is supported by core funding from the Wellcome Trust (203147/Z/16/Z).

The authors would like to thank all volunteers who took part in the study. We are grateful for the generous support of the imaging support team at UCL WCHN, particularly Daniel Bates, Chloe Carrick, Yasmin Feuozi, and Dimitra Moraiti for thoughtful discussions during scanning times. We would also like to thank Matthew Nour for helpful discussions during the setup of the study and for sharing analysis code. We thank Oliver Vikbladh and Neil Burgess for insightful discussions of task design & results and Leo Chi U Seak for helpful comments on an earlier version manuscript. We thank the members of the Max Planck UCL Centre for Computational Psychiatry and Ageing Research and the DeepMind Neuro-Lab for discussions at various stages of the project. We additionally thank Magda Dubois for creative suggestions of words describing the compound images used in the experimental task.

Contributions

LL, RM, ZK-N, RJD conceived and designed the study. LL and NC acquired the data. LL and NC analyzed the data. TE and SV advised on analyses. RM, ZK-N, RJD supervised analyses. LL and NC drafted the manuscript. All authors read and edited versions of the manuscript and approved the final version of the manuscript.

Conflict of Interest

ZK-N is employed by Google DeepMind. All other authors declare no conflict of interests.

References

Al Roumi, F., Marti, S., Wang, L., Amalric, M., & Dehaene, S. (2021). Mental compression of spatial sequences in human working memory using numerical and geometrical primitives. *Neuron*, *109*(16), 2627-2639.e4. https://doi.org/10.1016/j.neuron.2021.06.009

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(47), 29302–29310. https://doi.org/10.1073/pnas.1912341117

Baram, A. B., Muller, T. H., Nili, H., Garvert, M. M., & Behrens, T. E. J. (2021). Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. *Neuron*, *109*(4), 713-723.e7. https://doi.org/10.1016/j.neuron.2020.11.024

Barron, H. C., Dolan, R. J., & Behrens, T. E. J. (2013). Online evaluation of novel choices by simultaneous representation of multiple memories. *Nature Neuroscience*, *16*(10), 1492–1498. https://doi.org/10.1038/nn.3515

Barron, H. C., Garvert, M. M., & Behrens, T. E. J. (2016). Repetition suppression: a means to index neural representations using BOLD? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1705), 20150355. https://doi.org/10.1098/rstb.2015.0355

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, *183*(4), 954-967.e21. https://doi.org/10.1016/j.cell.2020.09.031

Collins, A. G. E., Ciullo, B., Frank, M. J., & Badre, D. (2017). Working memory load strengthens reward prediction errors. *Journal of Neuroscience*, *37*(16), 4332–4342. https://doi.org/10.1523/JNEUROSCI.2700-16.2017

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x

Conway, C. M., & Christiansen, M. H. (2006). *Statistical Learning Within and Between Modalities Pitting Abstract Against Stimulus-Specific Representations*.

Courellis, H. S., Minxha, J., Cardenas, A. R., Kimmel, D. L., Reed, C. M., Valiante, T. A., Salzman, C. D., Mamelak, A. N., Fusi, S., & Rutishauser, U. (2024). Abstract representations emerge in human hippocampal neurons during inference. *Nature*, *632*(8026), 841–849. https://doi.org/10.1038/s41586-024-07799-x

Dekker, R. B., Otto, F., & Summerfield, C. (2022). *Curriculum learning for human compositional generalization*. https://doi.org/10.1073/pnas

El-Gaby, M., Loyd Harris, A., R Whittington, J. C., Bhomick, A., Walton, M. E., Akam, T., & J Behrens, T. E. (2023). A Cellular Basis for Mapping Behavioural Structure. *BioArxiv*. https://doi.org/10.1101/2023.11.04.565609

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. https://doi.org/10.1016/j.neuron.2022.01.005

Flesch, T., Saxe, A., & Summerfield, C. (2023). Continual task learning in natural and artificial agents. In *Trends in Neurosciences* (Vol. 46, Issue 3, pp. 199–210). Elsevier Ltd. https://doi.org/10.1016/j.tins.2022.12.006

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *ELife*, *6*, 1–20. https://doi.org/10.7554/elife.17086

Garvert, M. M., Saanum, T., Schulz, E., Schuck, N. W., & Doeller, C. F. (2023). Hippocampal spatio-predictive cognitive maps adaptively guide reward generalization. *Nature Neuroscience*, *26*(4), 615–626. https://doi.org/10.1038/s41593-023-01283-x

Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *COGNITIVE SCIENCE*, *7*, 155–170.

Henin, S., Turk-Browne, N. B., Friedman, D., Liu, A., Dugan, P., Flinker, A., Doyle, W., Devinsky, O., & Melloni, L. (2021). Learning hierarchical sequence representations across human cortex and hippocampus. *Science Advances*, *7*(8), 1–13. https://doi.org/10.1126/sciadv.abc4530

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. In *Nature Neuroscience* (Vol. 19, Issue 3, pp. 404–413). Nature Publishing Group. https://doi.org/10.1038/nn.4238

Johnston, W. J., & Fusi, S. (2023). Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nature Communications*, *14*(1). https://doi.org/10.1038/s41467-023-36583-0

Kazanina, N., & Poeppel, D. (2023). The neural ingredients for a language of thought are available. In *Trends in Cognitive Sciences* (Vol. 27, Issue 11, pp. 996–1007). Elsevier Ltd. https://doi.org/10.1016/j.tics.2023.07.012

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(November), 1–28. https://doi.org/10.3389/neuro.06.004.2008

Kumar, S., Dasgupta, I., Marjieh, R., Daw, N. D., Cohen, J. D., & Griffiths, T. L. (2022). Disentangling Abstraction from Statistical Pattern Matching in Human and Machine Learning. *ArXiv*. http://arxiv.org/abs/2204.01437

Kurth-Nelson, Z., Behrens, T., Wayne, G., Miller, K., Luettgau, L., Dolan, R., Liu, Y., & Schwartenbeck, P. (2023). Replay and compositional computation. In *Neuron* (Vol. 111, Issue 4, pp. 454–469). Cell Press. https://doi.org/10.1016/j.neuron.2022.12.028

Lake, B. M., Salakhutdinov, R., & Tnenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*(2017). https://doi.org/10.1017/S0140525X16001837

Lehnert, L., Littman, M. L., & Frank, M. J. (2020). Reward-predictive representations generalize across tasks in reinforcement learning. *PLoS Computational Biology*, *16*(10), 1–27. https://doi.org/10.1371/journal.pcbi.1008317

Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. J. (2019). Human Replay Spontaneously Reorganizes Experience. *Cell*, *178*, 1–13. https://doi.org/10.1016/j.cell.2019.06.012

Luettgau, L., Erdmann, T., Veselic, S., Stachenfeld, K. L., Moran, R., Kurth-Nelson, Z., & Dolan, R. J. (2023). *Decomposing dynamical subprocesses for compositional generalization*.

Luettgau, L., Porcu, E., Tempelmann, C., & Jocham, G. (2022). Reinstatement of Cortical Outcome Representations during Higher-Order Learning. *Cerebral Cortex*, *32*(1), 93–109. https://doi.org/10.1093/cercor/bhab196

Luettgau, L., Tempelmann, C., Kaiser, L. F., & Jocham, G. (2020). Decisions bias future choices by modifying hippocampal associative memories. *Nature Communications*, *11*, 3318. https://doi.org/10.1101/802462

Mark, S., Moran, R., Parr, T., Kennerley, S. W., & Behrens, T. E. J. (2020). Transferring structural knowledge across cognitive maps in humans and models. *Nature Communications*, *11*(1), 1–12. https://doi.org/10.1038/s41467-020-18254-6

McElreath, R. (2020a). *rethinking: Statstical rethinking book package* (R package version 2.00).

McElreath, R. (2020b). *Statistical Rethinking* (2nd ed.). CRC Press. https://doi.org/10.1080/09332480.2017.1302722

Murphy, G. L. (1988). Comprehending Complex Concepts. *Cognitive Science*, *12*(4), 529–562. https://doi.org/10.1207/s15516709cog1204_2

Nieh, E. H., Schottdorf, M., Freeman, N. W., Low, R. J., Lewallen, S., Koay, S. A., Pinto, L., Gauthier, J. L., Brody, C. D., & Tank, D. W. (2021). Geometry of abstract learned knowledge in the hippocampus. *Nature*, *595*(7865), 80–84. https://doi.org/10.1038/s41586-021-03652-7

Nour, M. M., Liu, Y., Arumuham, A., Kurth-Nelson, Z., & Dolan, R. J. (2021). Impaired neural replay of inferred relationships in schizophrenia. *Cell*, *184*(16), 4315-4328.e17. https://doi.org/10.1016/j.cell.2021.06.012

O'Donnell, T. J., Goodman, N. D., & Tenenbaum, J. B. (2009). *Fragment Grammars: Exploring Computation and Reuse in Language*. www.csail.mit.edu

O'Donnell, T. J., Snedeker, J., Tenenbaum, J. B., & Goodman, N. D. (2011). Productivity and Reuse in Language. *Proceedings of Cognitive Science Conference*.

Pesnot Lerousseau, J., & Summerfield, C. (2024). Space as a scaffold for rotational generalisation of abstract concepts. *ELife*, *13*. https://doi.org/10.7554/eLife.93636

Reber, A. S. (1967). Implicit Learning of Artificial Grammars 1. In *JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR* (Vol. 6).

Redish, A. D. (2016). Vicarious trial and error. In *Nature Reviews Neuroscience* (Vol. 17, Issue 3, pp. 147–159). Nature Publishing Group. https://doi.org/10.1038/nrn.2015.30

Rmus, M., Ritz, H., Hunter, L. E., Bornstein, A. M., & Shenhav, A. (2022). Humans can navigate complex graph structures acquired during latent learning. *Cognition*, *225*. https://doi.org/10.1016/j.cognition.2022.105103

RStudioTeam. (2019). *RStudio: Integrated Development for R* (1.2.5033). RStudio, Inc. http://www.rstudio.com/

Rubino, V., Hamidi, M., Dayan, P., & Wu, C. M. (2023). Compositionality under time pressure. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*(4), 486–492. https://doi.org/10.1038/nn.3331

Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., Botvinick, M., Kurth-Nelson, Z., & Behrens, T. (2023). Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. *Cell*, *186*(22), 4885-4897.e14. https://doi.org/10.1016/j.cell.2023.09.004

Sheahan, H., Luyckx, F., Nelli, S., Teupe, C., & Summerfield, C. (2021). Neural state space alignment for magnitude generalization in humans and recurrent networks. *Neuron*, *109*(7), 1214-1226.e8. https://doi.org/10.1016/j.neuron.2021.02.004

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, *237*(4820). https://doi.org/10.1126/science.3629243

Sherman, B. E., Graves, K. N., & Turk-Browne, N. B. (2020). The prevalence and importance of statistical learning in human cognition and behavior. *Current Opinion in Behavioral Sciences*, *32*, 15–20. https://doi.org/10.1016/j.cobeha.2020.01.015

Sorscher, B., Ganguli, S., & Sompolinsky, H. (2022). Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas

Stan Development Team. (2020). *RStan: the R interface to Stan* (R package version 2.19.3.). http://mc-stan.org/

Story, G. W., Smith, R., Moutoussis, M., Isabel M. Berwian, I. M., Nolte, T., Bilek, E., Jenifer Z. Siegel, J. Z., & Dolan, R. J. (2024). A Social Inference Model of Idealization and Devaluation. *Psychological Review*, *131*(3), 749–780. https://doi.org/https://doi.org/10.1037/rev0000430

Tsividis, P. A., Loula, J., Burga, J., Foss, N., Campero, A., Pouncy, T., Gershman, S. J., & Tenenbaum, J. B. (2021). Human-Level Reinforcement Learning through Theory-Based Modeling, Exploration, and Planning. *ArXiv*, 1–67. http://arxiv.org/abs/2107.12544

Turk-Browne, N. B., Isola, P. J., Scholl, B. J., & Treat, T. A. (2008). Multidimensional Visual Statistical Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *34*(2), 399–407. https://doi.org/10.1037/0278-7393.34.2.399

Whittington, J. C. R., McCaffary, D., Bakermans, J. J. W., & Behrens, T. E. J. (2022). *How to build a cognitive map: insights from models of the hippocampal formation*. 1–25. http://arxiv.org/abs/2202.01682

Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. J. (2020). The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell*, *183*(5), 1249-1263.e23. https://doi.org/10.1016/j.cell.2020.10.024

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X. J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, *22*(2), 297–306. https://doi.org/10.1038/s41593-018-0310-2

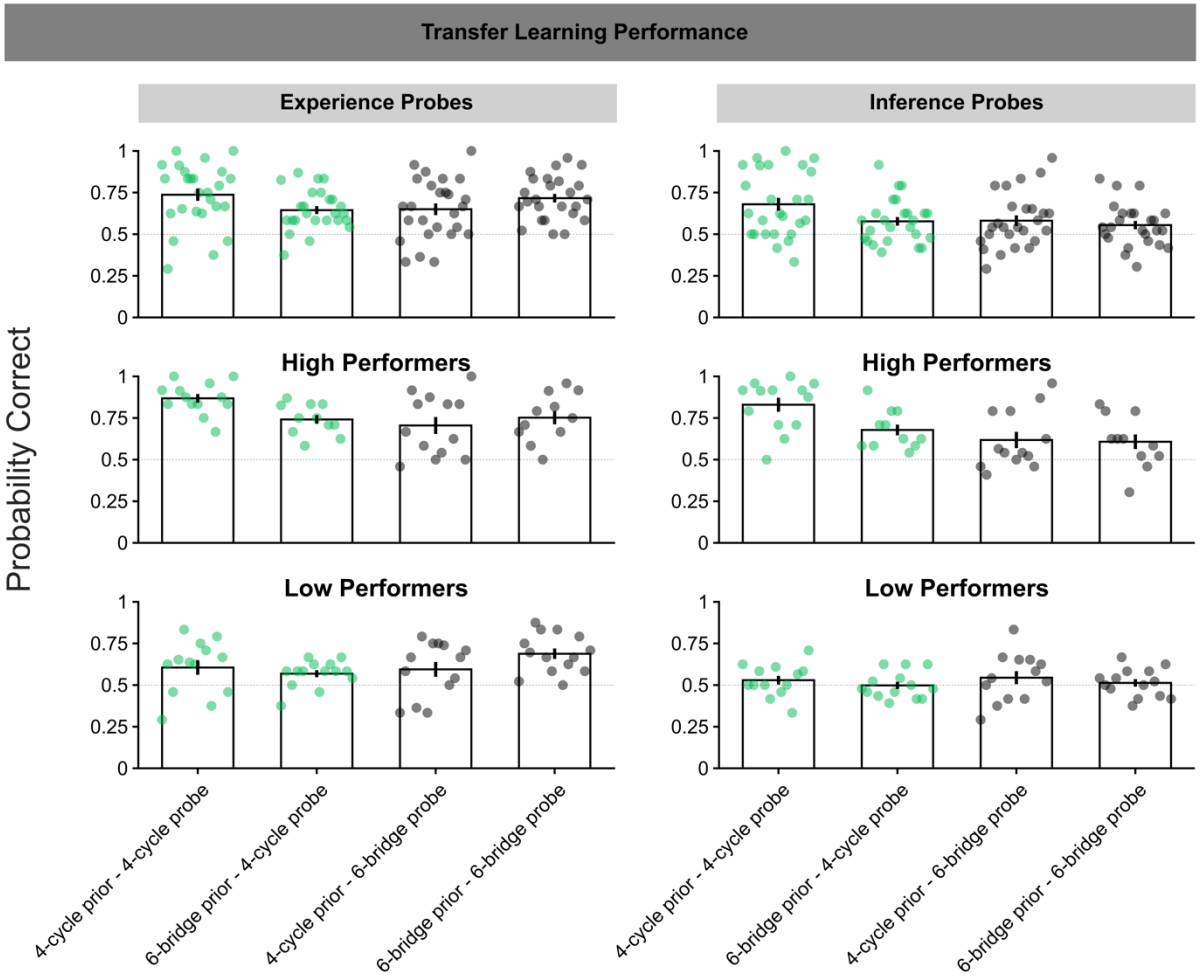
Zheng, X. Y., Hebart, M. N., Grill, F., Dolan, R. J., Doeller, C. F., Cools, R., & Garvert, M. M. (2024). Parallel cognitive maps for multiple knowledge structures in the hippocampal formation. *Cerebral Cortex*, *34*(2). https://doi.org/10.1093/cercor/bhad485

Zhou, J., Jia, C., Montesinos-Cartagena, M., Gardner, M. P. H., Zong, W., & Schoenbaum, G. (2021). Evolving schema representations in orbitofrontal ensembles during learning. *Nature*, *590*(7847), 606–611. https://doi.org/10.1038/s41586-020-03061-2

A neural mechanism for compositional generalization of structure in humans

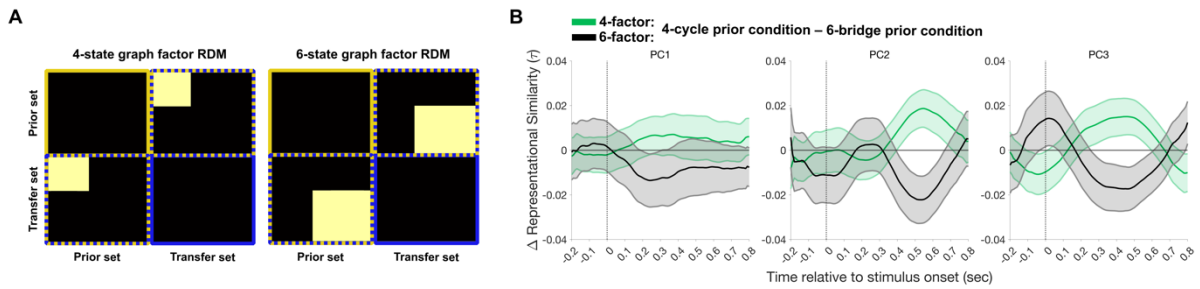
Supplementary Information

Supplementary Results



Supplementary Figure 1. Transfer learning performance.

Aggregated probabilities of correct answers during the transfer learning phase, separately for both experience (left) and inference probes (right) as a function of probed size (4-state cycle (green) vs 6-state bridge (black)) and condition (4-cycle prior vs 6-bridge prior). Full sample (top row), high performers (>50th percentile, middle row) and low performers (<=50th percentile, bottom row). Each dot represents one participant, bars represent the arithmetic mean of the distribution, error bars depict the standard error of the mean and the dashed gray line represents chance level point estimate (probability correct = 0.5).



Supplementary Figure 2. Representational similarity analysis, PC1-3

A) Theoretical representational dissimilarity matrices (RDM) regressed onto the empirical neural similarity matrices – separately for the 4-state graph factor features and the 6-state graph factor features. Lighter colors denote higher values. Note that the hypothesized similarity changes are only computed across the two feature sets used in prior learning (yellow) and transfer learning (blue), but not within each set, avoiding confounding task-related changes with feature similarity unrelated to abstract subprocess structure.

B) Time courses of condition differences (4-cycle prior condition – 6-bridge prior condition) on 4-state graph factor (green) and 6-state graph factor (black) representational similarity changes in PC1-PC3 from PRE to POST localizer. Shaded error bars indicate standard errors of difference. Horizontal line indicates a null effect.

Replay analyses

Building on work on human replay, as measured with MEG, in our and other research groups (e.g., Huang & Luo, 2024; Kern et al., 2024; Kurth-Nelson et al., 2016; Liu et al., 2019, 2021; Nour et al., 2021; Schwartenbeck et al., 2023) we tested for the presence of two qualitatively distinct types of sequential reactivation ("sequenceness") using temporally delayed linear modeling (TDLM). Specifically, we tested for different magnitudes of sequenceness that captured transitions between neural representations of compound stimuli ("compound sequenceness") and condition differences of sequenceness capturing transitions between neural representations of the individual features ("factorized sequenceness"). This analysis aimed at extending previous accounts of replay compositionality (Kurth-Nelson et al., 2023; Schwartenbeck et al., 2023). In these analyses we did not detect reliable, above permutation threshold sequenceness effects during a 5-minute rest period between prior and transfer learning, neither for compound stimuli nor for individual features. Such sequenceness effects were also absent during the initial 5-minute resting period before the PRE task localizer. Additionally, we did not observe any evidence for one-step transitions between neural representations of compounds or features at rest.

We also tested an additional hypothesis regarding a potential neural mechanism supporting gradual learning, and insight-dependent, decomposition of the experienced compound stimulus sequence into underlying subprocesses. Specifically, we examined for a trial-by-trial positive linear trend for factorized sequenceness and a trial-by-trial negative linear trend for compound sequenceness during the resting period between the experience and inference probe question (inter-probe intervals). Again, we did not detect significant and reliable sequenceness effects.

The absence of sequential reactivation findings may relate to several factors. Firstly, we We optimized hyper-parameters and classifier performance on visual localizer tasks and tested their generalization on memory reactivation during "no-task" resting periods. The L1-

regularization training procedure minimizes an overlap between representations of visually presented stimuli by reducing the magnitude of regression weights for non-influential sensors. Conceivably, this might impede an ability to reliably detect memory reactivation during “no-task” resting periods, rendering quantification of sequential memory reactivation challenging. Secondly, it is possible that the compositional generalization behavior seen in the present task may not rely on “offline” sequential reactivation of experienced task states. The behaviorally relevant information to be learned by participants encompassed one-step transitions between compound states (and features) alone, potentially rendering underpowered any sequential reactivation analyses that relies on multiple-step transitions of reactivated states.

References

Huang, Q., & Luo, H. (2024). Shared structure facilitates working memory of multiple sequences. *ELife*, *12*. https://doi.org/10.7554/eLife.93158

Kern, S., Nagel, J., Gerchen, M. F., Guersoy, C., Meyer-Lin-Denberg, A., Kirsch, P., Dolan, R. J., Gais, S., & Feld, G. B. (2024). Reactivation strength during cued recall is modulated by graph distance within cognitive maps. *ELife*. https://doi.org/10.7554/eLife.93357.3

Kurth-Nelson, Z., Behrens, T., Wayne, G., Miller, K., Luettgau, L., Dolan, R., Liu, Y., & Schwartenbeck, P. (2023). Replay and compositional computation. In *Neuron* (Vol. 111, Issue 4, pp. 454–469). Cell Press. https://doi.org/10.1016/j.neuron.2022.12.028

Kurth-Nelson, Z., Economides, M., Dolan, R. J., & Dayan, P. (2016). Fast Sequences of Non-spatial State Representations in Humans. *Neuron*, *91*(1), 194–204. https://doi.org/10.1016/j.neuron.2016.05.028

Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. J. (2019). Human Replay Spontaneously Reorganizes Experience. *Cell*, *178*, 1–13. https://doi.org/10.1016/j.cell.2019.06.012

Liu, Y., Mattar, M. G., Behrens, T. E. J., Daw, N. D., & Dolan, R. J. (2021). Experience replay is associated with efficient nonlocal learning. *Science*, *372*(6544). https://doi.org/10.1126/science.abf1357

Nour, M. M., Liu, Y., Arumuham, A., Kurth-Nelson, Z., & Dolan, R. J. (2021). Impaired neural replay of inferred relationships in schizophrenia. *Cell*, *184*(16), 4315-4328.e17. https://doi.org/10.1016/j.cell.2021.06.012

Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., Botvinick, M., Kurth-Nelson, Z., & Behrens, T. (2023). Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. *Cell*, *186*(22), 4885-4897.e14. https://doi.org/10.1016/j.cell.2023.09.004