Reviewed Preprint
v1 • July 28, 2025
Not revised

Reviewed Preprint
v2 • October 30, 2026
Revised by authors

A neural mechanism for compositional structure transfer in humans

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

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

For correspondence: lennart.luettgau@gmail.com

* These authors contributed equally

Competing interests: Work conducted while ZK-N was employed by Google DeepMind. All other authors declare no conflict of interests.

Reviewing editor: Roberto Bottini, University of Trento, Italy

© 2025, Luettgau et al. This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

eLife Assessment

This study provides **valuable** insights into humans' ability to generalize learned structural knowledge to new experiences that share underlying graph components but involve different stimuli. The behavioral and MEG evidence is **solid**, broadly supporting the transfer of abstract structural knowledge across contexts, although the findings speak more directly to the reuse of individual structural components than to their recombination. This study will be of interest to cognitive neuroscientists studying structure learning, abstract representations, and the generalization of knowledge across contexts.

https://doi.org/10.7554/eLife.107162.2.sa4

Abstract

A human ability to adapt to the dynamics of novel environments relies on abstracting and generalizing from past experiences. Previous research has focused on how humans generalize from isolated sequential processes, yet we know little about mechanisms that enable adaptation to more complex dynamics, including those that govern much everyday experience. Here, using a novel sequence learning task based on graph factorization, coupled with simultaneous magnetoencephalography (MEG) recordings, we ask whether reuse of experiential "building blocks" enable inference and generalization. Behavioral results were consistent with participants decomposing task experience into subprocesses, abstracting their dynamical structure away from their sensory specifics and transferring these to a new task environment. Neurally, this transfer was associated with a learning-induced, condition-specific, increase in neural similarity among stimuli conditional on their structural role in abstract subprocesses that were shared across task phases. Consistent with this, a complementary dynamical role decoding analysis revealed enhanced neural decodability for stimuli sharing identical structural transition types across experiential contexts, specifically when prior experience included the relevant graph structure. Decoding strength for these role representations was positively related to behavioral success in subprocess knowledge transfer, a relationship that will require future independent replication. These findings provide neural evidence consistent with the idea that a structural scaffolding mechanism supports generalization of experience to new contexts.

Introduction

A defining feature of human cognition is an exceptional ability to adapt rapidly to novel contexts. An influential 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 application of experiential knowledge to entirely new situations (Behrens et al., 2018; Lake et al., 2017; Mark et al., 2020; Shepard, 1987). Thus, much research has focused on 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). Human neural evidence indicates 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 little regarding how humans abstract from the more complex dynamics that characterize much everyday experience.

Computational modeling with artificial neural networks, trained to perform multiple tasks (Johnston & Fusi, 2023; Yang et al., 2019), and cross-species neural evidence, propose this may be achieved via a low dimensional neural coding of abstract, disentangled, task variables that enables a linear readout of relevant abstracted task dimensions (Bernardi et al., 2020; Courellis et al., 2024). Implicated brain structures include the 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; Courellis et al., 2024; Flesch et al., 2022, 2023; Zhou et al., 2021) and posterior parietal cortex (Flesch et al., 2022, 2023). Inference and generalization of knowledge could allow for a computationally efficient reuse or "recycling" of learnt representations in novel contexts akin to the abstraction of structural representations from sensory experiences proposed as implemented within the hippocampal-entorhinal system (Behrens et al., 2018; Whittington et al., 2020, 2022).

In previous 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 its ubiquity, we know little about the mechanisms of efficient adaptation to 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). Previously we found behavioral evidence for generalization of subprocesses that involved parsing complex task experience into component "building blocks". Based on computational modeling and behavioral evidence, we proposed a cognitive-computational mechanism whereby structural mapping of prior experience facilitates generalization of knowledge to help solve new tasks (Luettgau et al., 2024). While the latter data extended on previous behavioral findings regarding compositionality (Lake et al., 2015, 2017; Rubino et al., 2023), we still lack a neural account of how this might be implemented.

We hypothesized generalization relies on reuse of abstracted structural knowledge to form a representational scaffolding for new experiences. Throughout this work, we use "generalization" to refer to the ability to apply learned relational structure to novel stimulus compositions within the same task context, and "transfer" to refer to the reuse of (abstracted) representations across distinct task contexts or learning phases. Note we use "abstract" to refer to representations that are invariant to the sensory specifics of individual stimuli, capturing relational structure (e.g., the role a stimulus occupies within a graph) rather than surface features. Our task is structured compositionally in that every experienced compound stimulus is the joint product of two simultaneously evolving subprocesses, while the specific compound encountered during transfer learning is never observed during prior learning. We use 'decomposition' and 'transfer' to refer to the abstraction and reuse of a single such subprocess across learning phases, and reserve "compositional generalization" for the broader phenomenon whereby task performance depends

on appropriately decomposing and recombining structural knowledge, a capacity our design requires of participants. Here we acknowledge that our study speaks most directly to decomposition and transfer of individual structural components rather than to the act of recombination itself.

Under this framing, knowledge generalization involves mapping of inferred generative causes of an ongoing experience to support abstraction over concrete events (Mark et al., 2020; Pesnot Lerousseau & Summerfield, 2024; Whittington et al., 2020). We propose humans identify, neurally represent, and abstract the constitutive relational units of experience ("building blocks"). These may be reused as a structural scaffolding that biases and supports learning and inference about the relational units underlying new experiences, complementing rather than replacing direct experience. A structural scaffolding account of compositional generalization entails a prediction that, across different contexts, elements composing analogous subprocesses would show representational similarity, indicative of a shared abstract representation that enables transfer and reuse of experiential "building blocks".

While several studies have mapped the spatial location of low-dimensional neural representations of abstract and disentangled task variables to different brain areas – such as the hippocampal-entorhinal system (Baram et al., 2021; Bernardi et al., 2020; Courellis et al., 2024; Nieh et al., 2021), frontal cortex (Baram et al., 2021; Bernardi et al., 2020; Courellis et al., 2024; Flesch et al., 2022, 2023; Zhou et al., 2021) and posterior parietal cortex (Flesch et al., 2022, 2023). In parallel, M/EEG studies have shown that abstract, format-independent representations can emerge dynamically over time. For example, multivariate analyses of MEG data demonstrate cross-decoding of numerical magnitude across distinct symbolic formats (e.g., digits and dice), indicating a shared, format-invariant representation that unfolds with distinct temporal dynamics (Teichmann et al., 2018). More broadly, M/EEG evidence suggests that task-relevant information can be dynamically prioritized, compressed, and reformatted, indicative of a rapid temporal evolution of abstract and behaviorally relevant representations even when spatial localization remains coarse. Recent MEG data suggests that high-dimensional sensory inputs are progressively transformed into lower-dimensional, task-relevant, feature manifolds across distinct processing stages, reflecting dynamic gating and reformatting of information in support of flexible categorization (Duan et al., 2024). Complementary EEG–fMRI evidence also indicate abstract task goals are encoded in a low-dimensional neural geometry that emerges in frontal cortex and communicated to posterior regions to guide implementation of stimulus-specific representations, with the fidelity of this geometry predicting behavioral performance (Zhang & Yu, 2024). Likewise, converging behavioral and computational data indicates successful generalization depends on detecting and encoding abstract relational structure rather than stimulus surface features. In effect, training enables transfer when task phases share an underlying rule or abstract relationship, even across modalities, but not when an overlap is purely perceptual (Garner et al., 2016).

At the same time, reuse of shared abstract representations introduces a computational trade-off, such that overlapping codes that facilitate transfer can also induce interference, as observed in both humans and artificial neural networks (Holton et al., 2026). However, while there is now much work elucidating how learned representations are generalized, it remains unclear whether and how abstracted structural representations are expressed neurally during ongoing stimulus processing, and how they are shaped by prior structural experience. Thus, there is a need to characterize not only where abstract representations are located in the brain but also how they dynamically unfold and are reused across contexts in support of compositional generalization.

To address these questions, 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. Product graphs (Cartesian product), wherein 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 compelling and flexible test bed for probing compositional generalization of multiple (independent) and simultaneously unfolding subprocesses. Examples of such graph products and their factorizations, from a vast space of

possible combinations, are provided in Figure 1A. Here we take inspiration from prior studies in neuroscience (e.g., Barron et al., 2013; Rubino et al., 2023; Schwartenbeck et al., 2023) and a prior cognitive 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 ask three specific research questions. First, do stimuli that share the same abstract subprocess become more similar neurally following learning, and is this representational change specific to the subprocess experienced during prior learning? Second, do neural representations also come to reflect the fine-grained dynamical roles stimuli occupy within a subprocess? Third, does the strength of these neural representations relate to participants' ability to successfully transfer subprocess knowledge to new contexts? Addressing these questions allowed us to characterize the neural representational basis of compositional generalization, establishing when during stimulus processing abstract structural knowledge is expressed and how it is shaped by prior experience. 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" were represented similarly 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.

Results

In brief, participants (N = 50, N = 25 subjects per condition) performed a sequence learning task, observing image sequences produced by walks on product graphs constructed from smaller graph factors (Fig. 1A and B). The experiment consisted of two phases (prior learning and transfer learning, see task design details in Figure 1 and 2). During both task phases, participants viewed sequences of compound images comprising two elements, depicting meaningful scenes. The compound images used for the sequences were generated from combining images on two different graph factors (Fig. 1B), where these differed between-subjects and across the two task phases.

Using a between-subjects design, in a prior learning phase, 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, left, Fig. 1C for example sequences). 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, left). Each of these graph factors is considered a structural "building block" underlying participants' holistic experience.

The prior learning phase was followed by a transfer learning phase, identical for both conditions (Fig. 1D), but now compound image sequences were composed from entirely novel picture elements. The latter were the product of a 4-state cyclic graph factor and a 6-state bridge graph factor, respectively (Fig. 1B right). This between group design meant that, for each participant, the structure of only one of the two graph factors ("building blocks") that generated the compound image sequences was common across the prior and transfer learning phase (e.g., in the 4-state cycle prior condition, the 4-state cycle, but not the 6-state bridge, was present in both prior and transfer learning, Fig. 1B). Importantly, these graph factors were embedded by entirely different images across both task phases, rendering any learning unlikely based on stimulus similarity.

For brevity and clarity, we hereafter refer to the two prior learning conditions as Condition 4 (4-cycle prior) and Condition 6 (6-bridge prior), based on which graph structure each contributes to the transfer learning phase as a recurring structural component. Correspondingly, we refer to the 4-state graph factor (i.e., the 4-cycle) as Component 4 and the 6-state graph factor (i.e., the 6-state bridge) as Component 6. Under this notation, Condition 4 participants can reuse Component 4 knowledge during transfer, while Condition 6 participants can reuse Component 6 knowledge.

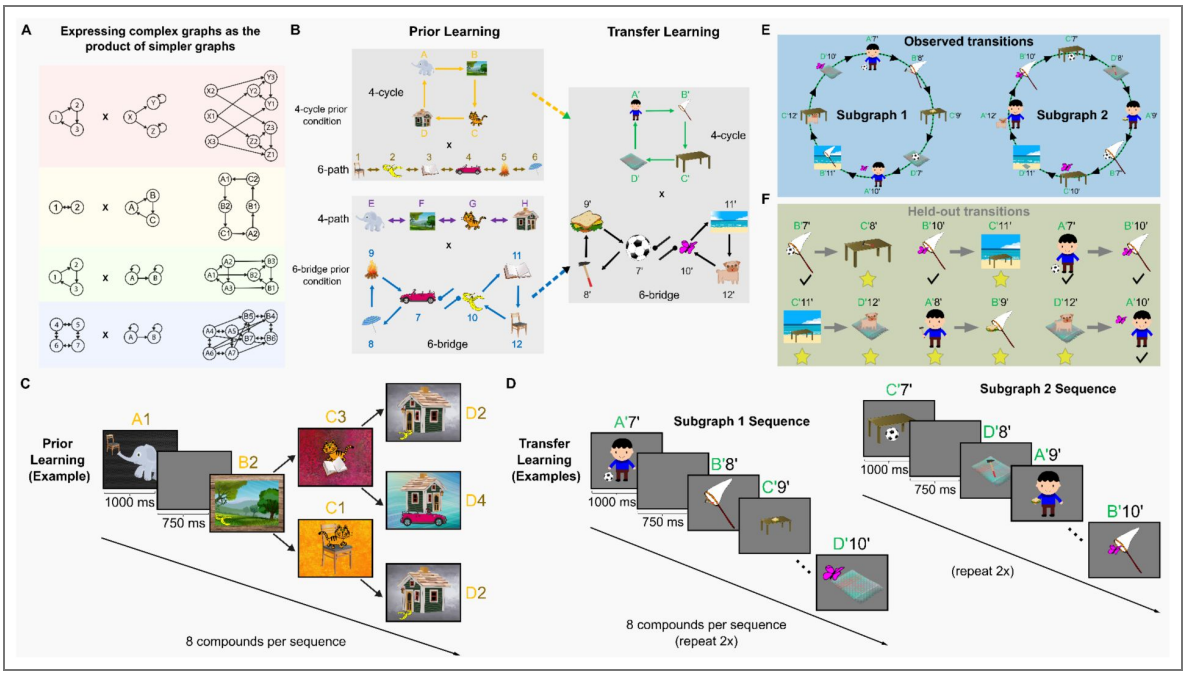


Figure 1. Experimental design

A) Exemplar product graphs (Cartesian product, right) selected from an 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 (both prior and transfer learning), graph factors were combined into compound graphs (product graphs) by merging their images. 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. Depicted is an example sequence generated by the product graph of the yellow graph factor and the gold graph factor used during prior learning (panel B, left, top). 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). After each sequence presentation (consisting of 8 compound stimuli), participants' knowledge of the graph structures was assessed in two ways, as detailed in Figure 2. **D)** During sequence presentations in transfer learning, participants only observed a subset of transitions from the compound graph formed by the green graph factor and the black graph factor, as shown in panel B (right), 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 F). This allowed us to test participants' understanding of the subprocesses generating the observed sequences. The same principles applied to prior learning. Depicted are 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. **E)** For illustration of the transfer learning compound graph, 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 instead had a "refractory" edge, such 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. **F)** Examples of held-out transitions between compound images used in the transfer learning phase. Note these entailed transitions between two compounds which were observed (as part of other transitions) during sequence presentations (denoted by a black tick below), or between an observed and a compound that had never been presented (denoted by yellow stars below) or between two new compounds. In any case, the specific transitions between compounds had never been observed during sequence learning.

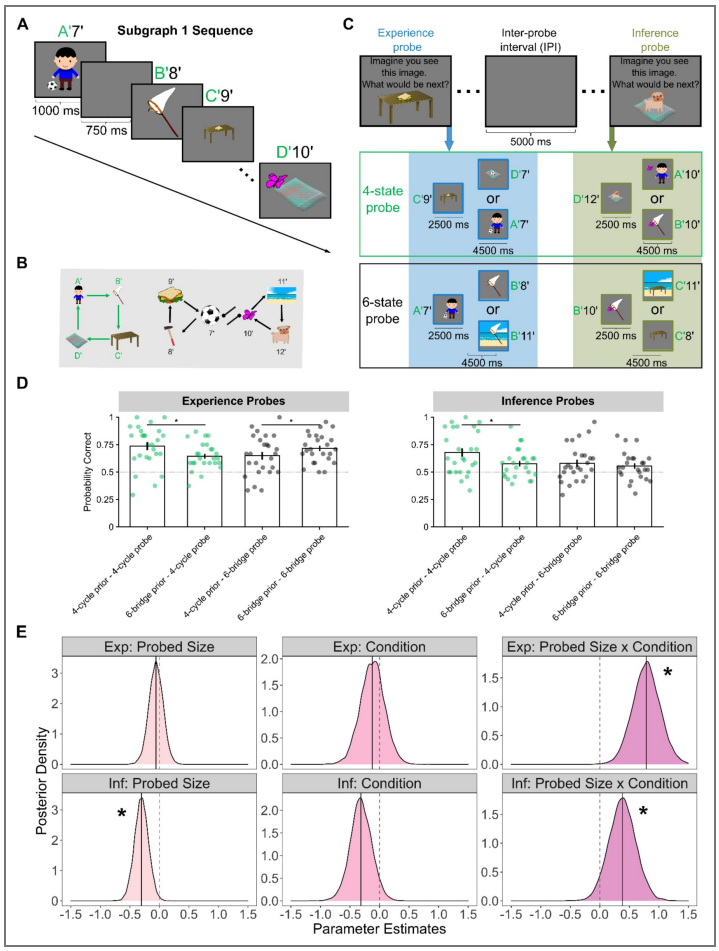


Figure 2. Behavioral task and transfer learning performance.

A) Example sequence produced by subgraph 1 of the product graph used in transfer learning (reproduced Fig. 1D left). 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 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 (Condition 4 vs Condition 6). 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). Asterisks indicate condition differences for which contrasts of posterior parameter estimates exclude 0. 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 (Component 4 or Component 6) and probe type (experience vs inference) effects, parameter estimates below zero indicate higher accuracy in Component 4 probes (vs. Component 6 probes) and experience probes (vs. inference probes), respectively. In the condition effect, parameter estimates below zero indicate higher accuracy in Condition 4 (vs. Condition 6). The positive interaction effect indicates selectively higher accuracy in Component 4 probes in Condition 4, and higher accuracy in Component 6 probes in Condition 6. Asterisks indicate GLM parameter estimates for which posterior distributions exclude 0.

Our central 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. To test this, in both task phases (i.e., prior and transfer learning), participants experienced sequences of a subset of all possible transitions between compound images (Fig. 1E and F illustrate this as an example for the transfer learning phase product graph). The remaining possible 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)). For details on the transitions, we refer to the Methods section.

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 (correct transition based on the generative graph structure) and a lure (incorrect transition) 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). Overall, the task allowed us to test if subjects utilize the structural "building blocks" underlying their experience (graph factors, subprocesses). Additionally, we could test whether knowledge of subprocess dynamics is reused compositionally in novel contexts. This is because in our design, despite using entirely different stimulus sets across the task phases, one of the abstract graph factors was consistent for participants in both conditions, allowing reuse of this specific structural knowledge. However, knowledge about the other graph factor that generated experiences in prior learning was irrelevant, enabling control for general learning effects.

In summary, participants viewed sequences of images that were combined from concurrent walks on two underlying graph structures ("building block"). In prior learning, they learned image sequences generated from two graph structures. In a subsequent transfer learning phase, participants saw new image sequences, but one of the two structural "building blocks" that underlie the observed image dynamics remained the same for each participant. Throughout both phases, participants were asked to predict upcoming images and infer never observed transitions. This task feature allowed us to test reuse of prior structural knowledge, which we predicted would reflect in higher accuracy in the consistently encountered "building block" (but not the other, irrelevant "building block") – providing evidence for a hypothesis that participants generalize components of learned structural dynamics despite changes in component sensory elements.

Abstracting and reusing subprocesses

A behavioral prediction was that individuals abstract experienced subprocesses (i.e., graph factors) from prior learning and reuse this knowledge during transfer learning. Thus, at transfer learning Condition 4 participants should perform better on Component 4 probes compared to Condition 6 participants, and vice versa for Component 6 probes. This predicts at transfer learning a disordinal interaction where the probability of being correct varies as a function of prior learning condition (Condition 4 vs Condition 6) and probed size (Component 4 vs Component 6) (Fig. 2D). To model candidate processes that might generate the observed transfer learning behavioral data, we defined four Bayesian multilevel GLMs (GLM1–4, 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). In post-hoc GLMs, 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)

Disentangling these effects using post-hoc contrasts [Condition 4 – Condition 6], showed that experience probes performance was higher for Component 4 probes in Condition 4 (vs Condition 6; Fig. 2D left panel, green dots; mean difference of 10,000 posterior samples: .36, CI = [.17; .56]) whereas for Component 6 probes, performance was superior in Condition 6 (vs Condition 4; Fig. 2D left panel, black dots; mean difference of 10,000 posterior samples: −.23, CI = [-.42; −.04]). In inference probes, performance on Component 4 probes was higher in Condition 4 (vs Condition 6; Fig. 2D right panel, green dots; mean difference of 10,000 posterior samples: .38, CI = [.20; .56]). For Component 6 probes, there was no evidence for a performance difference between conditions (Condition 4 – Condition 6; 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 findings replicate previous results from a larger cohort using a similar design (Luettgau et al., 2024). The Component 6 experience probe behavior utilized here indicates subjects not only extract and generalize subprocess dynamics for cyclical structures (as we previously showed), 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 are consistent with participants being influenced by experiential subprocess dynamics (Luettgau et al., 2024), where generative consistency across both task phases biases performance during transfer learning, though direct experience remains a dominant driver of behavior

Structural abstraction and transfer of subprocesses

A structural scaffolding account predicts elements of analogous subprocesses, across different experiential contexts, would show representational similarity reflective of adherence to a shared abstract representation, or neural scaffolding. Thus, this analysis addresses our primary research question as to whether abstract compositional representations are expressed neurally during stimulus processing, and whether their emergence is shaped by prior learning experience. To test the above hypothesis we recorded MEG signals during distinct localizer tasks, from before and after sequence learning (PRE and POST localizer task, Fig. 3A). We reasoned that a representational similarity in experiential “building blocks” should be reflected in a task-induced (PRE to POST) increase in neural similarity between distinct stimuli as a function of adherence to the same (compared to different) underlying relational structure, in this case adhering to identical subprocess dynamics (Fig. 3B). To test this, we used a feature similarity analysis (akin to 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 an initial step, to denoise sensor-level neural signals in a principled manner, we implemented a dimensionality reduction approach (principal components analysis, PCA, see Methods for details). Thus, for each participant, we implemented the following steps separately for PRE and POST localizer sessions. First, we extracted the minimal number of PCs that accounted for >=80% of the variance of neural activity across sensors for all trials and time points (−200 to 800 ms relative to stimulus onset). Next, for each time point, trial and PC we calculated the neural PC activity (i.e., PC score). By averaging PC activity across all trials presenting the same feature (i.e., element images

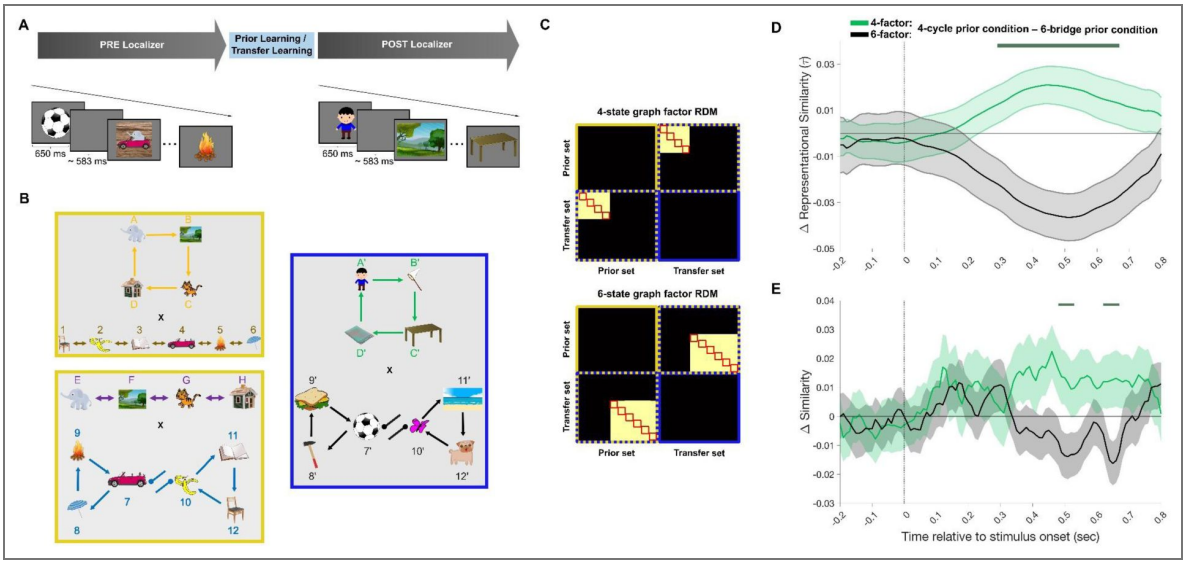


Figure 3. Neural dynamics of structural mapping between subprocesses.

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). Additionally, the PRE localizer also contained the compound images used during prior learning (not relevant for this analysis). A feature similarity analysis (akin to RSA) on stimulus-evoked neural activity recorded during the two localizer tasks (PRE and POST localizer) for the two feature sets, comprising 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 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. Red squares indicate structural mappings between stimulus sets (e.g., stimulus in the first, second, third position in the 4-state cycle), applied in E). **D)** Time courses of condition differences (Condition 4 – Condition 6) on Component 4 (green) and Component 6 (black) feature similarity changes in the weighted average PCs from PRE to POST localizer. Shaded error bars indicate standard errors of difference. Dark green line represents time points at which feature similarity changes for the condition x graph factor interaction effect were statistically significant, corrected for multiple comparisons at the cluster-level ($P_{FWE} < 0.05$, non-parametric permutation test). For prior learning group and graph factor main effects all *PFWE* > 0.118 (non-parametric permutation test). Horizontal line indicates a null effect. E) Time courses of condition differences (Condition 4 – Condition 6) in structural role–based representational similarity on Component 4 (green) and Component 6 (black), computed from sensor-level RSA in the localizer task (POST – PRE). Structural role similarity reflects increased neural similarity for stimuli occupying the same graph-theoretic role (e.g., boundary, intermediate, central) within each graph structure. Shaded error bars indicate standard errors of the difference. The dark green line denotes time points at which the prior condition × graph factor interaction was statistically significant, corrected for multiple comparisons at the cluster level ($P_{FWE} < 0.05$, non-parametric permutation test). For prior learning group and graph factor main effects all $P_{FWE} > 0.90$ (non-parametric permutation test). The horizontal line indicates a null effect.

f graph factors) we obtained an activity index for each PC, feature and time point. We next computed the dissimilarity between each pair of features as the absolute difference in their activity indices. This provided for a Representational Dissimilarity Matrix (RDM) at each time point, and for each PC. Next, for each time point, we calculated an across PC weighted average RDM (with weights being variance explained by each PC relative to the total variance explained by all PCs included in the analysis), resulting in a weighted average RDM for each participant, time point and session (PRE/POST). We then subtracted the POST weighted average RDM from the PRE weighted average RDM for each participant to obtain a PRE-POST difference weighted average RDM. Finally, to quantify feature similarity change at each time point, and for each participant, we computed Kendall's tau correlation coefficient between theoretical RDMs and PRE-POST differences in neural RDMs (similar to Luyckx et al., 2019; Nili et al., 2014). Our theoretical RDMs relate to a hypothesized abstract representation of subprocesses that reflect higher similarity between feature-pairs, where features belong to different learning phases (one from primary, the other from transfer) but share a structural graph component (separate design matrices for the 4-state graph factor and 6-state graph factor, Fig. 3C).

Within the PRE-POST change in weighted averaged RDMs, across participants we observed a qualitative pattern indicative of increased neural similarity between stimuli that shared the same underlying subprocess across task phases. A group-level GLM (including condition, graph factor main effects and a condition x graph factor effect) revealed that highly distinct stimuli that adhered to the same abstract relational structure (4-state cycle and 6-state bridge graphs, respectively), manifest a significantly greater neural similarity (Fig. 3D). For illustration purposes, we computed the average difference between feature similarity metrics for both prior learning conditions (4-cycle prior condition – 6-bridge prior condition), separately for each of the graph factors (see Fig. 3D). There was a statistically significant interaction effect of condition x graph factor spanning approximately 300 – 680 ms post-stimulus onset (cluster-level P_{FWE} = .004, corrected for multiple comparisons using non-parametric permutation tests, see Methods for details). The timing of this effect indicates that abstract structural information is expressed relatively early during stimulus processing, within a window broadly consistent with higher-order representational processing rather than early sensory encoding or late decision-related activity (Nour et al., 2021). Importantly, this time window should be interpreted as a marker of when abstract structural representations are neurally available, not as evidence for a discrete computational stage. We did not observe prior learning group and graph factor main effects (all P_{FWE} > 0.118, non-parametric permutation test). Post-hoc tests of condition differences for each graph factor indicated that within the time period where we observed a significant condition x graph factor effect, participants in Condition 4 showed higher positive change in feature similarity that mapped to Component 4 across both task phases, by comparison to participants in the Condition 6 (who experienced a 4-state path-graph in prior learning, Fig. 3D, green line, cluster-level P_{FWE}= .040, non-parametric permutation test), and vice versa for the Condition 6 (Fig. 3D, black line; dark green line in Fig. 3D represents time points at which feature similarity changes for both post-hoc contrasts are significant; cluster-level P_{FWE} = .002, non-parametric permutation test). Performing the former analysis with an unweighted average across the PCs that accounted for >= 80% of the variance (all PCs contribute equally) yielded highly similar results (interaction effect of condition x graph factor from 320 – 690 ms post-stimulus onset; cluster-level P_{FWE} = .007, corrected for multiple comparisons using non-parametric permutation tests). Importantly, to verify that the chosen PCA approach did not impose artificial structure within the data, or introduce other potential confounds, we also conducted the above analyses on sensor level MEG data. Highly similar results were obtained, with consistent results for different similarity metrics (absolute distance (L1 norm) and Euclidean distance (L2 norm), Supplementary Figure 3)), demonstrating that our results and conclusions are robust across analytic choices.

These results are not explained by participants simply learning to differentiate which of the 10 observed features pertain to Component 4 and 6, 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.

The observed increase in neural similarity between stimuli associated with the same graph factor could, in principle, reflect a more general categorical grouping of stimuli belonging to the same graph factor, rather than reflecting an abstract graph representation. To address this alternative account, we examined whether neural similarity in the localizer task is specifically structured according to the graph-theoretic role each stimulus occupies within its graph, using a finer-grained theoretical RDM that distinguishes between boundary, intermediate, and central positions. Structural roles were derived from each participant's learning task sequences and defined by graph topology. For Component 6, roles were assigned using a bridge-graph traversal template (for Condition 6, whose prior learning involved the bridge graph) or degree-based assignment (for Condition 4, whose prior learning involved a line graph): boundary nodes were defined as nodes with degree 1 (graph periphery); intermediate nodes as neighbours of boundary nodes; and central nodes as the remaining, more densely connected nodes. For Component 4, two roles were distinguished, boundary-like nodes (those encountered first and last in the sequence traversal) and central-like nodes (those encountered second and third), reflecting structural position within the graph rather than arbitrary temporal sequence. For further details of role assignment, see Methods. This analysis revealed a cluster-based permutation-corrected interaction ($P_{FWE} < 0.05$, non-parametric permutation test, Figure 3E) between prior learning group and graph factor at overlapping time windows with the previously reported PCA and sensor-level analyses investigating feature-mappings to abstract graph factors. We did not observe significant and multiple comparisons corrected prior learning group and graph factor main effects (all $P_{FWE} > 0.90$, non-parametric permutation test). These results indicate that representational similarity increased selectively for stimuli sharing the same structural role in the graph structure relevant to each group's prior experience. Critically, this effect was not observed uniformly for all stimuli associated with the same graph, arguing against a purely categorical grouping explanation. Instead, these findings provide evidence that learning induces representations that reflect higher-order relational structure, capturing graph-theoretic roles that are invariant across surface features of the embedding stimuli and support transfer across learning contexts.

Taken together, our findings indicate a task-induced, learning-related increase in neural similarity for stimuli sharing a structural role in subprocesses common across task phases and is compatible with a structural scaffolding mechanism, though we do not claim these representations are fully context-independent.

Abstract representation of dynamical roles in subprocesses

The previous analysis demonstrated that neural similarity increases for stimuli that share a common subprocess, consistent with a structural scaffolding account at a categorical level. Here we test a stronger and more specific prediction: that neural representations do not merely group stimuli by shared subprocess membership, but encode the precise relational, or "dynamical", role a stimulus occupies within that subprocess. Critically, this predicts enhanced neural similarity between transitions involving entirely different stimuli, stimuli that need not otherwise resemble one another, provided they occupy the same dynamical role within the graph structure. This goes beyond a generic categorical account, which predicts similarity based on subprocess co-membership alone, and instead requires that the specific relational position within the underlying structure is what is reused across contexts. Note this can only be ascertained for Component 6, as this graph alone incorporates stimuli in categorically distinct dynamical roles.

To test this, we focused on MEG data recorded during prior and transfer learning phases involving, in Component 6, compound stimulus presentations that adhered to specific abstract transition types (Figure 4A). Within Component 6, these transitions comprise three distinct types, reflective of their graph locations namely transitions between boundary and intermediate graph nodes (1), transitions between intermediate and central graph nodes (2) and transitions between central and central graph nodes (3) respectively (see Methods for details). During prior learning, we trained classifiers on neural activity from stimuli belonging to each of these three transition types. In the transfer learning phase, we tested this generalization of pre-trained classifiers to identical transition types, but where these transitions now entailed entirely different

stimuli. We used two-sample t-tests to ascertain classification accuracy between participant groups who had experienced different graphs during prior learning (Condition 4 experiencing a 6-path graph and Condition 6 experiencing a 6-bridge graph). This included multiple comparison correction based on nonparametric permutation tests.

In line with our hypothesis, following onset of a post-transition compound stimulus presentation, classification accuracy differences were evident for participants who had experienced different graphs during prior learning (6-path graph and 6-bridge graph), spanning two distinct temporal clusters between 135 and 350 ms post stimulus-onset (Figure 4B, first cluster (black bar): cluster-level corrected P_{FWE}= .035, two-tailed; second cluster (grey bar): P_{FWE} = .027, one-tailed, non-parametric permutation tests). Thus, enhanced neural decodability was evident for stimuli that shared transition types in the Component 6, but this effect was absent in participants who had experienced the 6-path graph (Component 4). This is consistent with emergence of neural similarity for stimuli that adhere to the same structural dynamics across experiential contexts, supporting a more fine-grained prediction of a structural scaffolding account.

To further test the robustness of the above effect, and to assess whether abstract dynamical roles are reflected in neural representational geometry during task performance, we conducted a complementary cross-phase representational similarity analysis. This analysis addresses directly whether neural representations of central nodes are stable across their distinct serial positions within a sequence, an expected property if representations reflect abstract structural identity rather than time-in-sequence. Specifically, for each participant we estimated neural patterns for each central node separately at its first and second occurrence (phase) within the sequence traversal, using occurrence-specific GLMs, and computed cross-phase similarity between prior and transfer learning phases. This approach controls for a potential time-in-sequence confound inherent in role-based comparisons, since now, the same functional node is compared across different serial positions (rather than comparing different stimuli with a common sequence position, by virtue of sharing role). Consistent with the decoding results, we observed increased cross-phase similarity in participants with prior experience of Component 6 relative to those with prior experience of Component 4 (Supplementary Figure 4). Specifically, a significant group difference was observed at a late stage of the trial (1.125–1.25 sec post-stimulus onset, P_{FWE} = 0.043, two-sample non-parametric permutation test, one-tailed), with additional differences at earlier timepoints that did not survive correction for multiple comparisons (0.125 sec: t = 2.18, p = .034; 0.79 sec: t = 2.84, p = .007). These effects indicate increased neural similarity for central node representations across phases in participants who experienced these roles consistently during prior learning. We consider these results complementary to the main analyses and interpret them with caution given that the group difference was only corrected for multiple comparisons at the late cluster. Importantly, the temporal profile of the cross-phase similarity effect does not perfectly overlap with the time windows identified in the original dynamical role decoding analysis, which limits the strength of the convergence argument. We therefore cannot fully rule out the possibility that the earlier decoding effects reported in Figure 4 partly reflect temporal position within the sequence rather than abstract structural role coding per se. We additionally note that this analysis is based on a smaller and less variable set of comparisons than either the original decoding analysis or the role-based RSA, since the bridge graph contains only two central nodes whose repeated occurrences can be directly compared in this way, limiting its statistical power in isolation. Nonetheless, the observation of a group-specific effect in the cross-phase analysis which directly controls for serial position provides support for the structural scaffolding account, and the late corrected cluster suggests that role-consistent representations do emerge across learning phases, albeit potentially at a later stage of processing than the decoding results alone would suggest.

While our data does not provide for an unequivocal interpretation of these dynamical roles, one possibility is that individuals differentiate transitions along Component 6 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

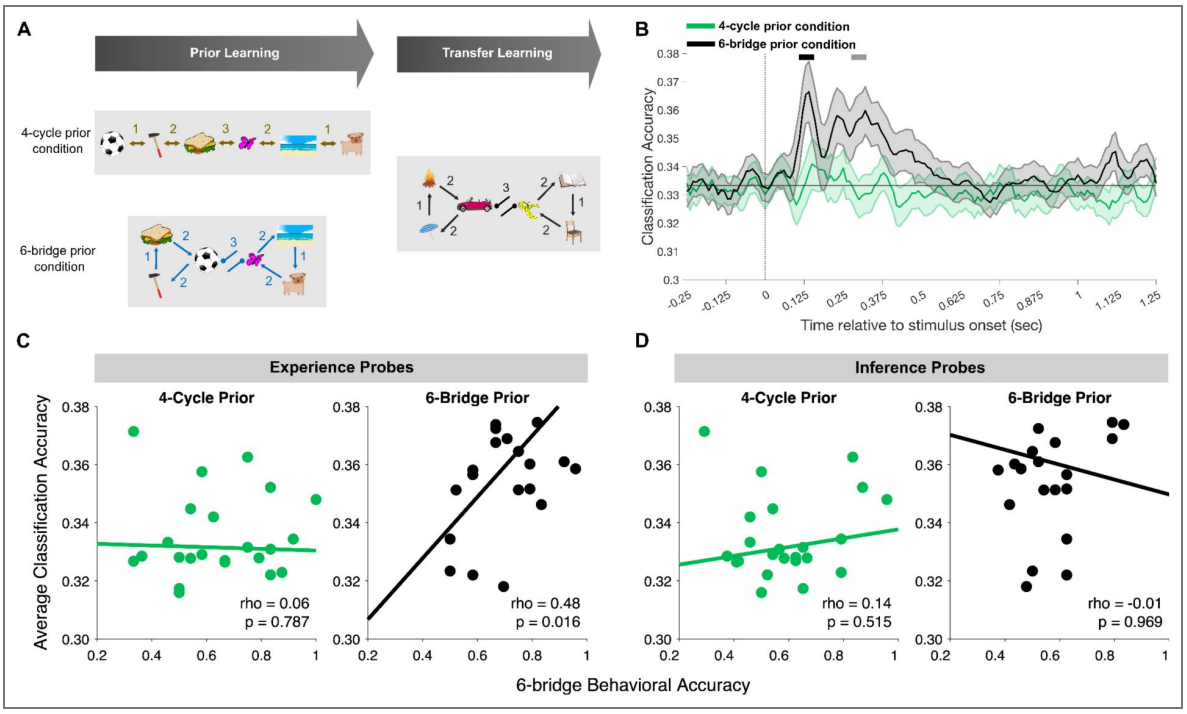


Figure 4. Abstract representation of subprocess dynamics.

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 and grey bars above indicate temporal clusters in which classification accuracy differences between conditions (6-bridge prior condition > 4-cycle prior condition) are significantly different (two-sample t-test), corrected for multiple comparisons at the cluster-level ($P_{FWE} < 0.05$, non-parametric permutation test (first cluster (black): $P_{FWE} = .035$, two-tailed; second cluster (grey): $P_{FWE} = .027$, one-tailed). Horizontal lines indicate chance level decoding accuracy (.33). C) Spearman correlations between average behavioral accuracy in Component 6 experience probes and per participant averaged classification accuracy across all timepoints post-image onset where there is a significant condition difference between Condition 4 and Condition 6 on classification accuracy (see panel B). Each datapoint represents an individual participant. Condition 6 (*rho* = .48, *p* = .016) showed a numerically stronger positive correlation than Condition 4 (*rho* = .06, *p* = .787, correlation difference: *Z* = 1.53, *p* = .127, Fisher *r* to *z*). **D)** Spearman correlations between average behavioral accuracy in Component 6 inference probes and per participant averaged classification accuracy across all timepoints post-image onset where there is a significant condition difference between Condition 4 and Condition 6 on classification accuracy (see panel B). Each datapoint represents an individual participant. There was no evidence for a correlation difference between Condition 6 (*rho* = −.01, *p* = .969) and Condition 4 (*rho* = .14, *p* = .515, correlation difference: *Z* = .48, *p* = .630, Fisher *r* to *z*).

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 found neural evidence consistent with dynamical role representations during transfer, we next asked whether the strength of these representations relates to participants' ability to successfully transfer subprocess knowledge to new contexts, a relationship that, if present, would provide converging behavioral support for the functional relevance of these representations, though we note this analysis is intended as exploratory and supportive rather than primary evidence. We thus investigated whether dynamical role classification accuracies (Fig. 4B) related to behavioral performance. Averaging classification accuracy over time points that showed significant condition differences (Fig. 4B), we found for experience probes a positive correlation between average classification accuracy and performance for Condition 6 (Condition 6: ρ = .48, p = .016, Spearman correlation; Condition 6: ρ = .06, p = .787; correlation difference: Z = 1.53, p = .127, Fisher r to z; Fig. 4C). No such relationship was evident for inference probes (Condition 6: ρ = –.01, p = .969, Spearman correlation; Condition 4: ρ = .14, p = .515; correlation difference: Z = .48, p = .630, Fisher r to z; Fig. 4D). These findings are consistent with, though not conclusive evidence for, dynamical role representations conveying behavioral benefits. Speculatively, such knowledge representation might enable individuals to constrain their predictions of 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". While we acknowledge that the correlation results appear inconsistent across experience and inference probes, it is important to consider that our MEG classifiers were trained and tested on experienced transitions alone, such that enhanced classification accuracy is more reflective of a neural representation of experienced transitions. Under this assumption, the MEG classifier would more closely align with behavioral accuracy for the experience probes (testing actually experienced transitions) than for the inference probes (testing never experienced but only implied transitions on the graph). See Discussion for further elaboration of this point, including why this asymmetry is expected under a model of gradual, imperfect structural learning.

The observed correlation is also consistent with a general learning-ability account, whereby higher-ability participants both better discover the Component 6 structure (that happens to be shared) across sessions and show higher behavioral accuracy overall. For example, higher-ability participants may rediscover this structure during transfer, without necessarily benefitting from having been exposed to it during primary learning. This ability-based account predicts that decoding strength should also relate to Component 4 accuracy, and that Component 4 and Component 6 accuracy should correlate within each prior learning group. Neither prediction was consistently supported (Supplementary Figure 5): the Component 4 decoding strength and Component 4 correlation was not significant (ρ = .38, p = .071), and did not differ significantly from the Component 4 decoding strength and Component 6 correlation (Steiger's test, p = .656); accuracy correlations between Component 4 and Component 6 were significant in only one prior learning group for each probe type, with no significant between-group differences (all p > .29).

Given this inconsistent pattern, we consider the general-ability account unsupported, though our low per-condition sample size limits power to rule it out. We therefore treat the brain-behavior correlation as selective to Component 6 experience probes in this sample, and treat it as supportive rather than primary evidence for a link between role representations and transfer, pending independent replication.

Finally, we tested whether sequential memory reactivation (replay) during rest periods could account for subprocess decomposition, using temporally delayed linear modelling (TDLM) of both compound and factorized sequenceness. We did not detect reliable above-permutation-threshold sequenceness effects during any rest period, for either compound stimuli or individual features (see Supplementary Results for full details and discussion of potential limiting factors).

Discussion

The philosophical adage of Heraclitus, "you can never step into the same river twice", reflects much everyday experience of the world. Yet even the most complex experiences tend to share structural similarity with the past, where this often entails a recurrence of experiential subprocesses. These recurrent world properties imply that exploiting regularities derived from past experience should enable agents to effortlessly master the relentless demands of complex and ever-changing environments (Gentner, 1983). More specifically, it has been 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., 2024; Pesnot Lerousseau & Summerfield, 2024). Indeed, growing evidence consistent with a compositional (re-)use of acquired knowledge provides a plausible cognitive mechanism to explain how humans transfer knowledge from prior experiences (Barron et al., 2013; Lake et al., 2015, 2017; Luettgau et al., 2024). Here, we replicate previous behavioral findings for parsing of subprocesses that generate the dynamics of experience, replicating (Luettgau et al., 2024), and now provide neural evidence consistent with a proposed abstract representational scaffolding, including providing neural evidence for representational changes consistent with a structural scaffolding mechanism. The neural realization of such a scaffolding is the subject of computational accounts of hippocampal and entorhinal cortex function, where grid cell activity is proposed as providing a basis set representation of abstract structure inferred from sensory experiences (Behrens et al., 2018; Mark et al., 2020; Whittington et al., 2020, 2022). However, the evolution of such structure mappings have remained elusive. We hypothesized that compositional generalization in our task is partly achieved by abstracting relational units of prior experience ('building blocks') that bias and facilitate learning in new contexts, complementing rather than supplanting direct experience. In principle, this could be enabled by mapping new incoming sensory information onto pre-existing structural templates. We conducted three empirical tests of this "structural scaffolding" hypothesis. Addressing our first research question, whether stimuli sharing the same abstract subprocess come to be represented more similarly following learning, and whether this is specific to the subprocess experienced during prior learning, we show task-induced, condition-specific changes in neural similarity of unrelated stimuli, presented across different task phases and embedded in novel complex dynamics, conditional on these adhering to identical subprocesses and mapping onto specific structural roles. Specifically, we observed a condition-specific, learning-induced, increase in neural similarity between stimuli mapping to the same abstract structural subprocesses around 300 – 680 ms post-stimulus onset, a finding consistent with structured temporal relationships shaping neural representations across contexts (Baram et al., 2021), and with a modest degree of task structure generalization. Here, we extend these findings by detailing the associated neural representations supporting compositional generalization. The representational change signal observed during stimulus processing (peaking around 500 ms post-stimulus onset), echoes previous findings showing emergence of an abstract structural representation purported to account for aberrant inference and delusions in psychopathology (Nour et al., 2021). While this similarity change is consistent with neural state space alignment (Sheahan et al., 2021), it is also predicted by a more parsimonious categorical grouping/labeling of different stimuli belonging to the same abstract structural subprocess.

Secondly, addressing our second research question, whether neural representations reflect the fine-grained dynamical roles stimuli occupy within a subprocess, we found that stimuli sharing the same dynamical role across different learned structures come to exhibit neural similarity, consistent with a more specific prediction of the structural scaffolding account that goes beyond generic subprocess grouping. Decoding strength for these role representations showed a positive, selective correlation with experience probe behavioral performance for the structurally relevant graph. While the direction of this brain-behavior correlation is consistent with a structural scaffolding account, the strength of this relationship relative to an alternative graph (4-cycle) did not reach statistical significance, and we caution this result needs independent replication to ascertain its robustness and specificity. We acknowledge that such a brain-behavioral correlation was not present for inference probes. However, we caution that the MEG classifiers were trained

and tested on experienced transitions alone – the classification accuracy thus reflects a neural representation of experienced transitions. Considering this, we reason that the MEG classification signal would be expected to more closely align with the behavioral accuracy in the experience probes (testing actually experienced transitions), instead of inference probes (testing never experienced but only implied transitions on the graph).

Addressing our third research question, whether the strength of identified neural representations relate to behavioral transfer, we found a positive relationship between decoding strength and experience probe performance that was numerically larger for the structurally relevant (6-bridge) graph than for the irrelevant (4-cycle) graph, though this difference did not reach significance; no corresponding relationship was evident for inference probes. Thus, a behavioral accuracy group difference in predicting successive states for the 6-bridge inference probe performance was of negligible magnitude and did not align with the robust effect observed for experience probes (Fig. 2D), where the latter replicates and extends a previous finding using a 6-state cycle graph and predicted by computational simulations using successor feature models (Luettgau et al., 2024). We speculate that the disparity between probe types is likely due to experience probes and inference probes engaging dissociable cognitive processes – the former relying on predictive representations acquired from experience, the latter relying on an ability to perform mental operations on learned representations of independent feature transitions. This may increase the likelihood of errors due to complexities of the inferred state spaces. However, the observed differences could also relate to previously reported effects of explicit knowledge on performance levels (Luettgau et al., 2024) or be driven by capacity limitations in working memory taxing larger subprocesses more strongly (Collins et al., 2017; Collins & Frank, 2012). It is also likely that structural knowledge is not acquired uniformly across compounds. Because abstract roles are learned gradually and with uncertainty, role inference is inherently context-dependent: the role of a given stimulus is not inferred in isolation, but in relation to the compound within which it appears. When one element of a compound is already known with high confidence, the role of the second element can be inferred more reliably; when both elements are themselves uncertain, inference becomes less precise. Learning precision may therefore vary across stimulus pairings depending on their frequency, co-occurrence statistics, and position within a sequence, and we would expect corresponding variability in participants' ability to infer dynamical roles, and hence to predict outcomes, across different compound types. This implies that an MEG signal, which indexes role-based representational structure specifically for experienced compounds, most directly reflects knowledge precision for those items. A correlation with inference-probe accuracy requires an additional inferential step, from experienced to novel compounds, introducing noise proportional to the uncertainty in the learned representations. In principle, better structural knowledge about experienced compounds should support better generalisation to novel ones, but we might be insufficiently powered to detect this weaker relationship. We therefore interpret an absence of a significant correlation for inference probes not as inconsistent with a structural scaffolding account, and suggest that a null result here should be interpreted with caution rather than as evidence against it. More broadly, an absence of a strong correlation between aggregate neural decoding measures and aggregate inference-probe performance should not itself be taken as evidence against abstract learning. Inference performance is likely supported by multiple sources of information, including partial structural knowledge and heuristic strategies, whereas neural decoding captures the stability of representational geometry at the population level. The localizer RSA results reported above demonstrate that abstract role representations are present in the neural signal, even though their task-dependent expression may not always be directly mirrored in inference-probe behavior.

At the very least our findings indicate striking learning, generalization and transfer abilities based on factorization of complex experiences into underlying structural elements, parsing these into distinct subprocesses derived from past experience, and forming a representation of the dynamical roles these features play within distinct subprocesses. These representations, in turn, support prediction of successive features leading to more accurate performance.

Our neural findings suggest that building and maintenance of representations of subprocesses allows factoring out the idiosyncrasies of the complex dynamics of experience. While the principal aim of the present study was to characterize a neural mechanism for compositional generalization of structural knowledge, it is important to consider that MEG recordings reflect an aggregate neural signal. Consequently, our analyses do not support inferences about regional brain sources of the observed signal. A prior literature indicates this type of mapping relates to activity of entorhinal cortex grid cells, where these compute invariant representations of relational structure underlying experience that is abstracted away from its 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 organizing information within structured representations, including abstract neural geometries, facilitates few-shot concept learning (Sorscher et al., 2022). The neural signatures we identify are consistent with the emergence of low dimensional representations of abstract 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, evidence indicates that 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) and that 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 encodes structure across experiential contexts. Within this account, similar state transition dynamics and dynamical roles encountered in distinct experiential contexts, results in stimuli that adhere to these being represented closer in a neural activity space (Bernardi et al., 2020; Courellis et al., 2024; Sheahan et al., 2021), putatively constraining a hypothesis space over abstract structural components that make up new experiences.

An important avenue for future research relates to the neural mechanism that supports an algorithmic implementation of the proposed “structural scaffolding” and how subprocess dynamics are disentangled. One plausible mechanism involves a separation of subprocess representations via preferential reactivation at different positions in the theta cycle, or by preferred reactivation coupled to distinct peaks or troughs in theta oscillations (analogous to a role of theta oscillations in vicarious trial-and-error computations, Redish, 2016). Our findings hint at such a mechanism by showing successor predictive representations are encoded neurally during transfer learning, finding decoding evidence of successor features from a current state’s neural activity pattern (Supplementary Figure 2 and Supplementary Results). Alternatively, gradual learning, and insight-dependent decomposition of an experienced compound stimulus sequence into underlying subprocesses could be enabled by human replay, extending previous accounts that propose a role in compositionality (Kurth-Nelson et al., 2023; Schwartenbeck et al., 2023). We tested this directly using TDLM-based sequenceness analyses during rest periods but did not detect reliable replay effects within the constraints of our experimental design (see Supplementary Results for full details), leaving this as an open question for future work. Relatedly, there is the question of how, and under what conditions, the structure of previous experiences map to current experience. We concede that the present experiment examines generalization across two families of graph structures, rather than across a broader space of possible graphs. While we consider it unlikely that our findings are idiosyncratic to cycles, lines and bridge-like graphs, we acknowledge that further work will be needed to support abstraction beyond the limited graph structures studied here. 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 also contribute to pinpointing the precise functional anatomy, including 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 that it provides an empirical test of an abstract structural scaffolding as a mechanism supporting decomposition and transfer of structural knowledge across learning contexts. We argue that such a mechanism enables reuse of previously abstracted relational units that biases learning and shapes neural representations in novel contexts, within a task whose compositional structure requires participants to integrate this reused knowledge with entirely new structural components, though our evidence speaks most directly to the transfer of individual components rather than to their combination.

Our findings characterize the neural representational basis of this proposed mechanism, establishing that abstract structural knowledge is expressed during ongoing stimulus processing and is shaped by prior structural experience. In the latter account, similar state transition dynamics encountered in highly distinct experiential contexts are represented more closely in neural activity space than dissimilar ones, allowing a sharing of components across distinct experiences. Finally, we note that failures in efficient knowledge decomposition, involving acquisition and reuse of an abstract neural scaffolding, might impact a flexibility of cognition that is 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 (±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., 2024) subjects were presented with sequences of compound images. Compound images were constructed by overlaying two independently varying visual features drawn from separate feature sets. Each feature set corresponded to a single graph factor, such that the state of each graph factor independently determines one feature of the compound image. For example, a compound image in state (A, 3) would display feature A from the cyclic factor and feature 3 from the path factor, superimposed or juxtaposed into a single image.

The set of all compound images and their transitions was defined by the Cartesian product of two graph factors: transitions between compound images were valid if, and only if, one factor alone changed state according to its own transition rules while the other factor remained fixed.

Observed sequences of image compositions composed from 10 individual features. Each sequence consisted of 8 compound images with 12 compound images in total across sequences. In a between-subjects design, during prior learning one group of participants (condition 1, 4-state cycle

prior) observed sequences of image compositions 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). For instance, in condition 1, a valid sequence might proceed: (A,1) → (A,2) → (A,3) → (B,3) → (C,3) → (C,4) → (D,4) → (D,5), where each step reflects a transition in either the cyclic (letter) or path (number) factor, but never both simultaneously.

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 transitions between compound images from two disjoint subgraphs of the compound graph (8 unique compound images per subgraph), composed from 10 novel individual features, Fig. 1C) 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). 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. 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. Because sequences were drawn from a subset of the product graph, only a portion of all 24 possible compound images (and their transitions) were encountered directly. Transitions between compound images could thus be: (i) directly experienced in a sequence, (ii) never presented in a sequence but implied by the factorization structure, or (iii) transitions between compound images from different subgraphs, which were never valid. The held-out set of transitions described in (ii) and (iii) 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 14^{th} or 12^{th} (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 ($R^{'} \approx 1.00$) and by visually inspecting Markov chain convergence using trace- and rank-plots. For all models fitted we found $R^{'} = 1.00$ 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, 2024). 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.

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

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

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

$$\mu_\alpha \sim Normal(0,1)$$
$$\mu_{PT} \sim Normal(0,1)$$
$$\sigma_{subject} \sim Half-Normal(0,1)$$
$$\sigma_{PT} \sim Half-Normal(0,1)$$
$$R \sim LKJcorr(2)$$

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

$$Correct_{i,j,k,n} \sim Binomial(number\ of\ probes_{i,j,k,n}, p_{i,j,k,n})$$
$$logit(p_{i,j,k,n}) = \mu_\alpha + \mu_{SIZE[i]} + \mu_{PT[j]} + \mu_{COND[k]} + \alpha_{subject[n],SIZE[i]} + \gamma_{subject[n],PT[j]} + \mu_{SIZEx\ COND}$$

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

$$\begin{bmatrix}\gamma_{n,1}\\ \gamma_{n,2}\end{bmatrix} \sim MVNormal\left(\begin{bmatrix}0\\0\end{bmatrix}, S_{PT}\right)$$
$$S_{SIZE} = \begin{pmatrix}\sigma_{subject} & 0\\ 0 & \sigma_{SIZE}\end{pmatrix} R_{SIZE} \begin{pmatrix}\sigma_{subject} & 0\\ 0 & \sigma_{SIZE}\end{pmatrix}$$
$$S_{PT} = \begin{pmatrix}\sigma_{subject} & 0\\ 0 & \sigma_{PT}\end{pmatrix} R_{PT} \begin{pmatrix}\sigma_{subject} & 0\\ 0 & \sigma_{PT}\end{pmatrix} \quad \textit{(Eq. 3)}$$

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

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

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

$$logit\left(p_{i,j,k,n}\right) = \mu_{\alpha} + \mu_{SIZE[i]} + \mu_{PT[j]} + \mu_{COND[k]} + \alpha_{subject[n],SIZE[i]} + \gamma_{subject[n],PT[j]} + \mu_{SIZEx\ COND} + \mu_{SIZExPTxCOND} \quad \textit{(Eq. 4)}$$

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

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

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

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

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

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 (± standard error): GLM4 = 1015.2 (±21.25), GLM3 = 1020.2 (±22.11)). All other GLMs explained the data less well than GLM4 (WAIC values (± standard error): GLM2 = 1101.4 (±33.04), GLM1 = 1151.9 (±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 across all participants due to excessive baseline noise. Preprocessing was conducted separately for each recording block, identical to previous studies from our laboratory (e.g., 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). On average across all MEG runs and participants, 8.67 sensors (SD = 7.36) and 9.13 trials (SD = 11.46) were excluded (using default osl_detect_artefacts settings; channel significance level for Generalised Extreme Studentised Deviate = 0.05). We did not apply MaxFilter as our data were acquired on a CTF system, for which MaxFiltering is not designed. Sensor interpolation was not performed as bad channels are

excluded prior to ICA decomposition following standard OSL pipeline practice; given that excluded sensors represented on average ~3% of the total 273 sensors, we do not expect this to have meaningfully affected the results. 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 (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. Note that for each participant different MEG sensors were excluded, based on the identified noise level. Only sensors that were not identified as noisy consistently across localizer or main task runs – depending on the respective analyses – were included. We treated noisy sensors to be missing completely at random.

MEG data analysis

Feature similarity analysis

We conducted a feature similarity analysis (inspired by 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, 20 PCs per subject, explaining >= 90% of the variance). 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 [feature x 1] vector of regression weights, as an estimate of the unique PC activation related to each feature. We repeated this procedure for all PCs. We then calculated the pairwise absolute differences between features as distance metric at each time point, for each PC separately. This generated a symmetrical [20 × 20] Representational Dissimilarity Matrix (RDM) at each time point. To retain as much stimulus similarity information contained within the neural data as possible, while discarding noisy components, we used the 20 PC-related RDMs to compute a weighted average RDM for each participant at each time point (in both localizer tasks separately). The weighting of the average RDM was based on explained variance of the PCs included in the analysis. For each participant individually this encompassed the top RDM computed on the PCs that cumulatively explained >= 80% of the variance of the data. To enable an estimate of task-induced PRE-POST changes, we then computed the difference in weighted averaged RDMs (POST minus PRE). 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 feature 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 feature similarity change was smoothed over time with a 25 ms Gaussian kernel prior to computing Kendall's tau.

In addition to the PCA-based feature analysis, we performed a sensor-level RSA. At each time point, we estimated condition-specific activation patterns using a GLM applied separately to each sensor (dummy-coded stimulus identity). RDMs were then computed across stimuli using mean absolute difference (L1 norm) as the primary distance metric. For completeness, we additionally evaluated Euclidean distances (L2 norm), which yielded qualitatively similar results (Supplementary Figure 3).

To test whether localizer representations of abstract graph structure generalised across sets experienced in the main task, we performed RSA separately for PRE and POST localisers. Structural roles were derived from each participant's learning-phase transition sequences. For 4-state graphs, nodes were assigned to two matched role bins (boundary-like vs central-like) based on first-encounter order within each phase. For 6-state graphs, roles were determined either from node degree (6-line: boundary, intermediate, central). For 6-bridge graphs we used a predefined structural template. Specifically, the first eight valid transitions in the learning phase were mapped onto a fixed sequence of structural roles (central–intermediate–boundary–central–central–intermediate–boundary–central). Each node was assigned a role based on the majority role it occupied across its occurrences in these eight steps (ties resolved in favour of central > intermediate > boundary). For the localiser data, at each time point and sensor we estimated condition-specific activations using a GLM with 20 dummy-coded regressors (10 identities × 2 sets). This yielded sensor-wise beta patterns for each condition. Cross-set representational dissimilarities were computed using an L1 distance metric (mean absolute difference across sensors), resulting in 4×4 and 6×6 cross-set RDMs at each time point. We then summarised each RDM according to structural role: within-role distances (same role across sets), between-role distances (different roles), and their difference (delta = between - within). To control for unequal pair counts, between-role comparisons were subsampled to match the number of within-role pairs (200 repetitions). This produced time-resolved structural generalisation indices for 4-state and 6-state graphs, which were submitted to group-level statistical analyses as described above.

We used nonparametric tests to identify timepoints where there was evidence for a condition (4-cycle prior condition and 6-bridge prior condition) x graph factor (4-factor and 6-factor) interaction effect, based on a group-level GLM (including condition, graph factor main effects and a condition x graph factor effect), correcting for multiple comparisons over time. At each timepoint, we computed the GLM and extracted the *F*-value for the interaction effect. We then identified contiguous clusters of time points where the observed *F*-values exceeded a predefined threshold (critical *F*-value) and computed the sum of *F*-values (*F*-sum) within each cluster. To correct for multiple comparisons, we performed a cluster-based permutation procedure that involved 1000 permutations. In each permutation, we shuffled the condition labels across participants before recomputing the *F*-values for the interaction effect at each time point, extracting the largest cluster-level *F*-sum from each permutation to build an empirical null distribution. An empirical *p*-value was computed by calculating the proportion of permutations for which the empirical absolute *F*-sum was smaller than or equal to the absolute permuted *F*-sums. A cluster in the observed data was deemed significant family-wise error corrected at the cluster-level if the empirical *p*-value was $P_{FWE} < 0.05$. Similarly, we conducted permutation tests for the post-hoc tests, comparing the two conditions on the two graph factors (two-sample tests) using the same cluster-based correction approach.

Decoding of dynamical roles

The following analysis aimed to assess neural evidence for an abstract representation of transitions in the 6-state graph factor.

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 4A). 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.

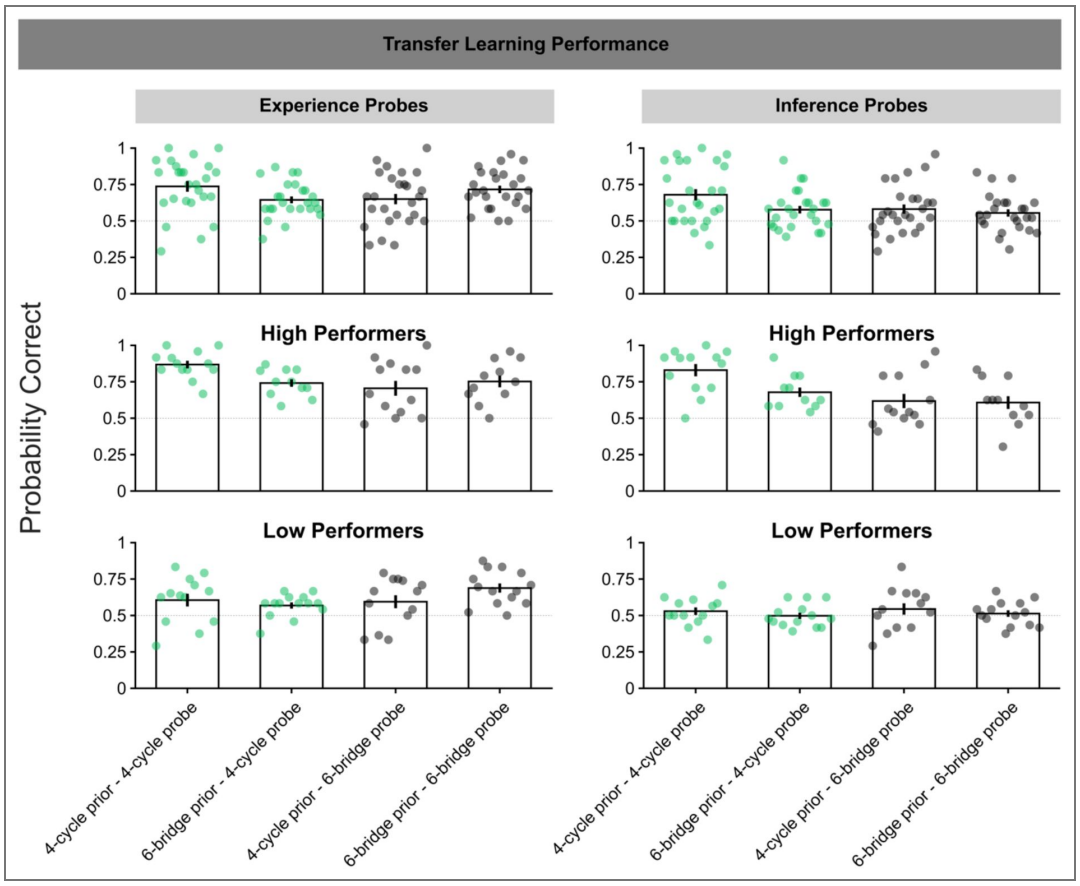
We analyzed neural activity in a window from 250 ms pre- to 1250 ms post-stimulus onset for each post-transition stimulus. The 1250 ms post-stimulus period encompassed the 1000 ms image display duration. Neural activity during stimulus presentations following each transition label was used to train L1-regularized 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.

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, similarly as described for the RSA, we performed a cluster-based permutation procedure with 1000 permutations. At each timepoint, we computed a two-sample t-test and extracted the *t*-value. We then identified contiguous clusters of time points where the observed *t*-values exceeded a predefined threshold (critical *t*-value) and computed the sum of *t*-values within each cluster. In each permutation, we shuffled the condition labels across participants before recomputing the *t*-values for the condition comparison at each time point. We then identified contiguous clusters of time points where the observed *t*-values exceeded a predefined threshold and used the largest cluster-level *t*-sum from each permutation to build an empirical null distribution. An empirical *p*-value was computed by calculating the proportion of permutations for which the empirical absolute *t*-sum was smaller than or equal to the absolute permuted *t*-sums. A cluster in the observed data was deemed significant family-wise error corrected at the cluster-level if the empirical *p*-value was $P_{\mathrm{FWE}} < 0.05$.

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).

Supplementary Results



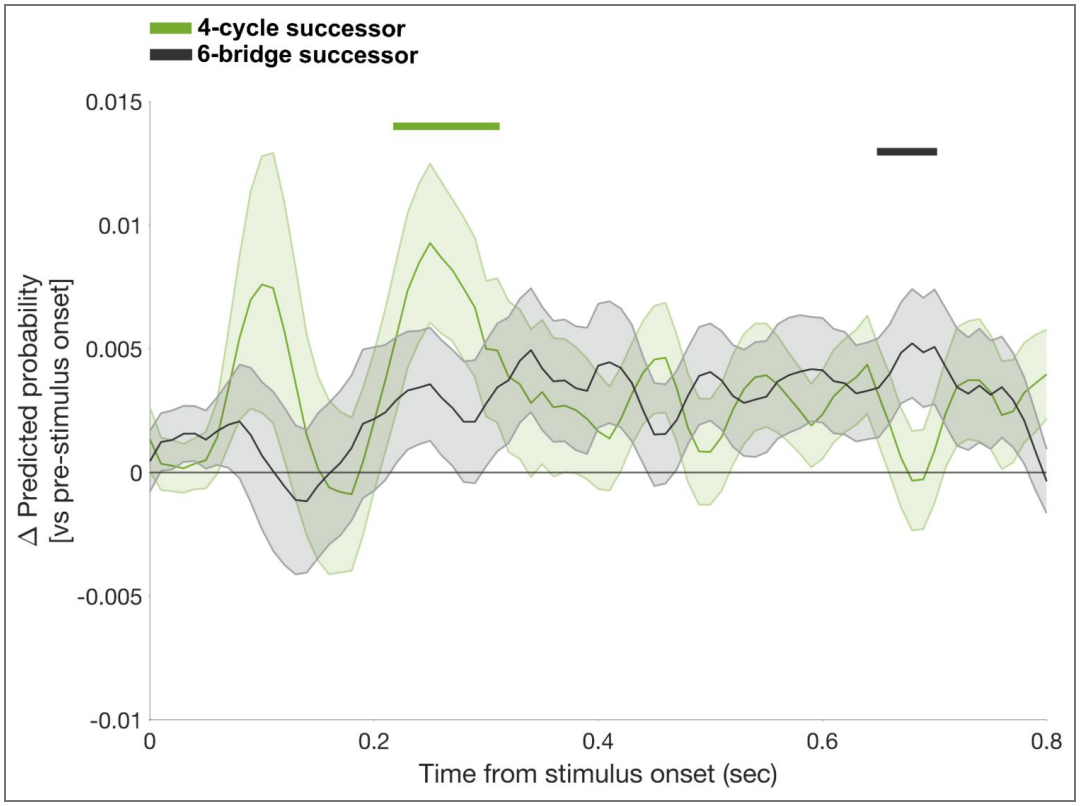
Supplementary Figure 1. Transfer learning performance. Aggregated probabilities of correct answers during the transfer learning phase, separately for both experience (left) and inference probes (right) as a function of probed size (Component 4 (green) vs Component 6 (black)) and condition (Condition 4 vs Condition 6). 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).

Neural successor feature representations

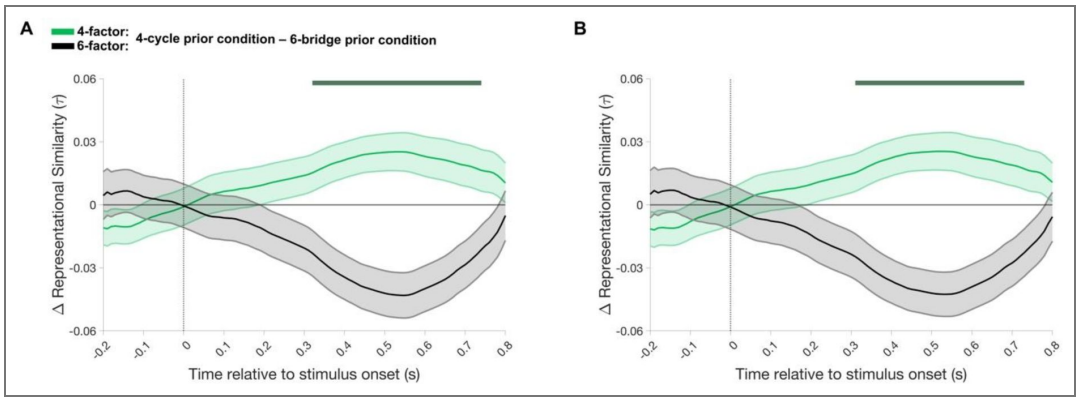
Increased neural similarity for stimuli belonging to identical abstract subprocesses supports a knowledge transfer of relevant structure across experiential contexts. On this basis, we asked about neural dynamics of successor feature representations.

We focused on characterizing neural signals that reflect knowledge representation of experienced subprocess dynamics during transfer learning. If indeed 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 trained classifiers on neural activity recorded for each individually presented feature during the PRE task localizer phase (Fig. 3A, see Supplementary 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 Component 4 and Component 6 successor features (see Methods for details on the classification scheme). Across several time points post-stimulus onset, we found increased predicted probabilities that were above average pre-stimulus onset predicted probabilities (Supplementary Fig. 2, $N = 49$ due to convergence issues of classifiers in one participant in Condition 4) for successor features

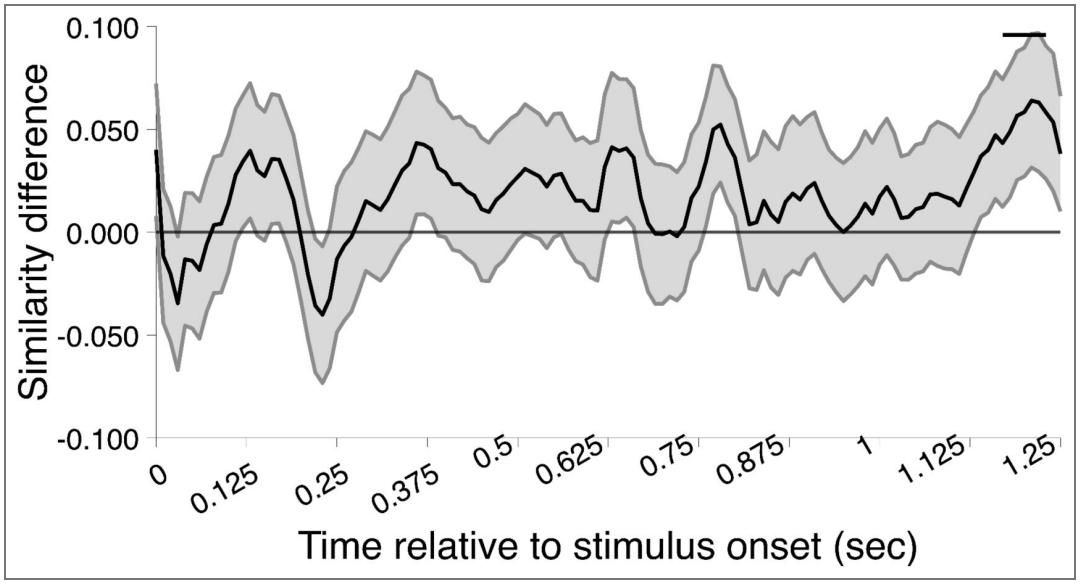
pertaining to both graph factors (cluster-level corrected *P*FWE= .002 and *P*FWE= .037, non-parametric permutation test; for Component 4 and Component 6 successor features, respectively), consistent with a neural representation of successor features for both experienced subprocesses after stimulus onset.



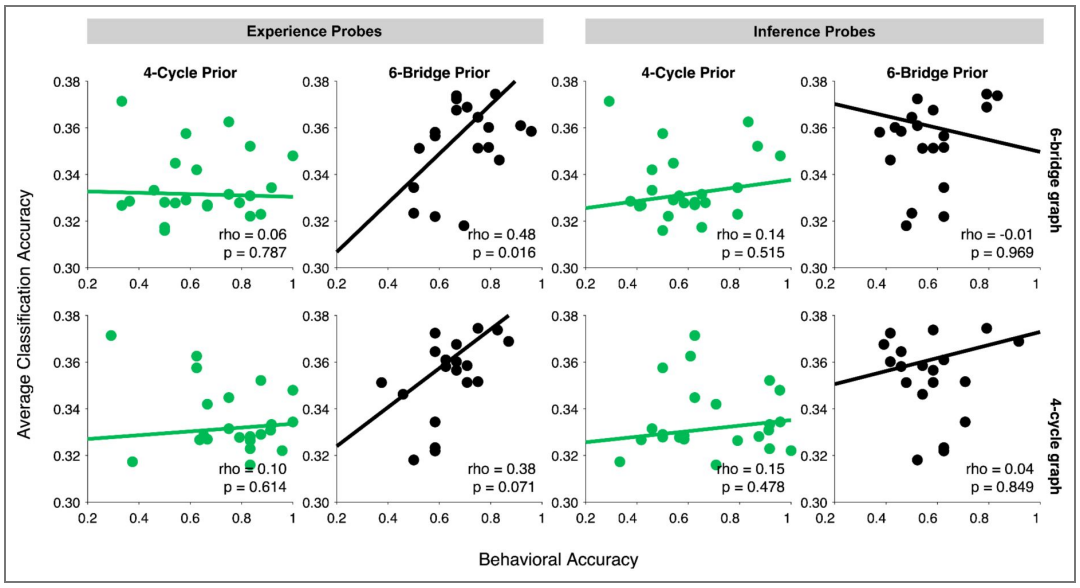
Supplementary Figure 2. Temporal dynamics of successor state predictions. Time courses of change in predicted probability of classifiers (reactivation) relative to average pre-stimulus onset baseline during transfer learning across both conditions, related to all stimulus presentations, for the respective successor features of 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 Component 4 and Component 6 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 on top index temporal clusters wherein the predicted probability is statistically significant above 0, corrected for multiple comparisons at the cluster-level ($P_{FWE} < 0.05$, non-parametric permutation test, one-tailed).



Supplementary Figure 3. Alternative distance metrics for results shown in Figure 3D. A) Time courses of condition differences (4-cycle prior condition – 6-bridge prior condition) on Component 4 (green) and Component 6 (black) feature similarity changes in the weighted average PCs from PRE to POST localizer at the sensor-level, using absolute distance (L1 norm) as similarity metric. Shaded error bars indicate standard errors of difference. Dark green line represents time points at which feature similarity changes for the condition x graph factor interaction effect were statistically significant, corrected for multiple comparisons at the cluster-level ($P_{FWE} < 0.05$, non-parametric permutation test). Horizontal line indicates a null effect. B) Same as A) for Euclidean distance (L2 norm) as similarity metric.



Supplementary Figure 4. Role-based representational similarity. Time courses of cross-phase representational similarity for central nodes of Component 6, computed from sensor-level data. Time courses reflect the difference in L1 distance between Condition 4 and Condition 4 group participants (positive values indicate higher similarity for Condition 6). For each participant, central node neural patterns were estimated separately for the first and second presentation of each central node within a given sequence. Similarity was computed as the mean L1 distance across four cross-phase pairs: central node A (first presentation in prior, second in transfer), central node A (second presentation in prior, first in transfer), and the equivalent pairs for central node B. Timecourses are baseline-corrected (pre-stimulus period) and restricted to post-stimulus timepoints. The shaded error bar indicates the standard error of the difference between groups. The black bar above the time course indicates timepoints where Condition 6 differed significantly from Component 4 (corrected for multiple comparisons at the cluster level, $P_{FWE} < 0.05$, two-sample non-parametric permutation test, one-tailed).



Supplementary Figure 5. Brain–behavior correlations with Component 4 and Component 6 performance during transfer learning. Scatter plots show the relationship between average neural decoding strength, computed at timepoints exhibiting a significant group difference (Fig. 5B), and behavioral transfer-learning performance, separately for Condition 4 (green) and Condition 6 (black). Rows show performance on Component 6 (top) and Component 4 (bottom). Columns show experience probe performance (left two columns) and inference probe performance (right two columns). Each point represents one participant; solid lines show least-squares fits. Rho and p-values (bottom right of each panel) reflect Spearman correlations with permutation-based significance testing (10,000 permutations, two-tailed).

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 in these data may relate to several factors. Firstly, 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.

Supplementary Methods

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 Component 4 and Component 6 subprocess successor feature representation. We then identified contiguous clusters of time points where the observed *t*-values exceeded a predefined threshold (critical *t*-value) and computed the sum of *t*-values within each cluster. To correct for multiple comparisons, we performed a cluster-based permutation procedure with 1000 permutations. In each permutation, we shuffled the sign of each of the successor feature activations before recomputing the t-values vs 0 at each time point. We then identified contiguous clusters of time points where the observed *t*-values exceeded a predefined threshold and used the largest cluster-level *t*-sum from each permutation to build an empirical null distribution.

An empirical *p*-value was computed by counting the number of times the original cluster sum was greater than or equal to the permuted *t*-sums. A cluster in the observed data was deemed significant family-wise error corrected at the cluster-level if the empirical *p*-value was *P*FWE <

0.05.

Cross-phase representational similarity analysis

Complementary to the decoding analysis in Figure 4, we conducted a cross-phase representational similarity analysis (RSA) to test whether neural representations of central nodes were stable across their distinct serial positions within a sequence, a property expected under abstract structural role coding but not under serial position coding. We focused on central nodes of the 6-state graph, as these nodes appear twice within each sequence traversal of the bridge graph: the left central node occurs at steps 1 and 4, the right central node occurs at steps 5 and 8, allowing first and second presentations to be dissociated.

Structural roles were assigned separately for prior and transfer learning phases. For line graphs (experienced by the Condition 4 during prior learning), central nodes were identified from node degree in the adjacency matrix (boundary: degree 1; intermediate: neighbors of boundary; central: remaining nodes), and the two central nodes were ordered by first appearance in the trial list. For bridge graphs, central nodes were identified from the predefined 8-step traversal template.

At each time point and sensor, condition-specific activations were estimated using occurrence-specific GLMs: one GLM fitted to trials corresponding to the first occurrence of each number in the trial list, and one fitted to trials corresponding to the second occurrence. Component 6-specific beta estimates were extracted from each GLM separately, yielding four sets of neural patterns per subject: prior-phase first occurrence, prior-phase second occurrence, transfer-phase first occurrence, and transfer-phase second occurrence.

Cross-phase representational dissimilarity was then computed using an L1 distance metric (mean absolute difference across sensors) across four cross-presentation pairs: Left central (prior first, transfer second), left central (prior second, transfer first), right central (prior first, transfer second), and right central (prior second, transfer first). These pairings compare the same node across different serial positions and different learning phases, directly controlling for time-in-sequence as a potential confound. The mean L1 distance across the four pairs constituted the cross-presentation dissimilarity index at each time point, where lower values indicate greater neural similarity. Timecourses were baseline-corrected by subtracting the pre-stimulus mean (timepoints 1–25, corresponding to −250 to 0 ms) per participant, and analyses were restricted to post-stimulus timepoints. Group differences between Condition 6 and Condition 4 participants were assessed using a two-sample non-parametric cluster permutation test (cluster-forming threshold $p = 0.10$, cluster-level alpha = 0.05, 1,000 permutations, one-tailed, consistent with the directional prediction that Condition 6 participants would show greater cross-phase neural similarity for central nodes). The results are presented in Supplementary Figure 4.

Data availability

Behavioral and MEG data and the code 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). Open access funding provided by Max Planck Society.

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.

Additional information

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.

Author ORCID iDs

Lennart Luettgau: https://orcid.org/0000-0001-9370-9606
Zeb Kurth-Nelson: https://orcid.org/0000-0002-0119-1289
Rani Moran: https://orcid.org/0000-0002-7641-2402
Raymond J Dolan: https://orcid.org/0000-0001-9356-761X

Note

This reviewed preprint has been updated to update the email correspondence.

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**:2627-2639.e4. https://doi.org/10.1016/j.neuron.2021.06.009 | PubMed

Allen KR, Smith KA, Tenenbaum JB (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**:29302-29310 https://doi.org/10.1073/pnas.1912341117 | PubMed

Baram AB, Muller TH, Nili H, Garvert MM, Behrens TEJ (2021) Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. *Neuron* **109**:713-723.e7. https://doi.org/10.1016/j.neuron.2020.11.024 | PubMed

Barron HC, Dolan RJ, Behrens TEJ (2013) Online evaluation of novel choices by simultaneous representation of multiple memories. *Nature Neuroscience* **16**:1492-1498 https://doi.org/10.1038/nn.3515 | PubMed

Barron HC, Garvert MM, Behrens TEJ (2016) Repetition suppression: a means to index neural representations using BOLD?. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**:20150355 https://doi.org/10.1098/rstb.2015.0355 | PubMed

Behrens TEJ, Muller TH, Whittington JCR, Mark S, Baram AB, Stachenfeld KL, Kurth-Nelson Z (2018) What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron* **100**:490-509 https://doi.org/10.1016/j.neuron.2018.10.002 | PubMed

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

Collins AGE, Ciullo B, Frank MJ, Badre D (2017) Working memory load strengthens reward prediction errors. *Journal of Neuroscience* **37**:4332-4342 https://doi.org/10.1523/JNEUROSCI.2700-16.2017 | PubMed

Collins AGE, Frank MJ (2012) How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience* **35**:1024-1035 https://doi.org/10.1111/j.1460-9568.2011.07980.x | PubMed

Conway CM, Christiansen MH (2006) Statistical Learning Within and Between Modalities Pitting Abstract Against Stimulus-Specific Representations. *Psychological Science* **17** https://doi.org/10.1111/j.1467-9280.2006.01801.x

Courellis HS, Minxha J, Cardenas AR, Kimmel DL, Reed CM, Valiante TA, Salzman CD, Mamelak AN, Fusi S, Rutishauser U (2024) Abstract representations emerge in human hippocampal neurons during inference. *Nature* **632**:841-849 https://doi.org/10.1038/s41586-024-07799-x | PubMed

Dekker RB, Otto F, Summerfield C (2022) Curriculum learning for human compositional generalization. *PNAS* **119**:e22055821 https://doi.org/10.1073/pnas.2205582119

Duan Y, Zhan J, Gross J, Ince RA, Schyns PG (2024) Pre-frontal cortex guides dimension-reducing transformations in the occipito-ventral pathway for categorization behaviors. *Current Biology* **34**:3392-3404 https://doi.org/10.1016/j.cub.2024.06.050 | PubMed

El-Gaby M, Loyd Harris A, R Whittington JC, Bhomick A, Walton ME, Akam T, J Behrens TE (2023) A Cellular Basis for Mapping Behavioural Structure. *bioRxiv* https://doi.org/10.1101/2023.11.04.565609 | PubMed

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**:1258-1270.e11. https://doi.org/10.1016/j.neuron.2022.01.005 | PubMed

Flesch T, Saxe A, Summerfield C (2023) Continual task learning in natural and artificial agents. *Trends in Neurosciences* **46**:199-210 https://doi.org/10.1016/j.tins.2022.12.006 | PubMed

Garner K, Lynch CR, Dux PE (2016) Transfer of training benefits requires rules we cannot see (or hear). *Journal of Experimental Psychology: Human Perception and Performance* **42**:1148 https://doi.org/10.1037/xhp0000215 | PubMed

Garvert MM, Dolan RJ, Behrens TE (2017) A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife* **6**:1-20 https://doi.org/10.7554/elife.17086 | PubMed

Garvert MM, Saanum T, Schulz E, Schuck NW, Doeller CF (2023) Hippocampal spatio-predictive cognitive maps adaptively guide reward generalization. *Nature Neuroscience* **26**:615-626 https://doi.org/10.1038/s41593-023-01283-x | PubMed

Henin S, Turk-Browne NB, 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**:1-13 https://doi.org/10.1126/sciadv.abc4530 | PubMed

Holton E, Braun L, Thompson J, Grohn J, Summerfield C (2026) Humans and neural networks show similar patterns of transfer and interference during continual learning. *Nature Human Behaviour* **10**:111-125 https://doi.org/10.1038/s41562-025-02318-y | PubMed

Huys QJM, Maia TV, Frank MJ (2016) Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience* **19**:404-413 https://doi.org/10.1038/nn.4238 | PubMed

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

Kazanina N, Poeppel D (2023) The neural ingredients for a language of thought are available. *Trends in Cognitive Sciences* **27**:996-1007 https://doi.org/10.1016/j.tics.2023.07.012 | PubMed

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

Kumar S, Dasgupta I, Marjieh R, Daw ND, Cohen JD, Griffiths TL (2022) Disentangling Abstraction from Statistical Pattern Matching in Human and Machine Learning. *ArXiv* https://doi.org/10.48550/arxiv.2204.01437 | PubMed

Kurth-Nelson Z, Behrens T, Wayne G, Miller K, Luettgau L, Dolan R, Liu Y, Schwartenbeck P (2023) Replay and compositional computation. *Neuron* **111**:454-469 https://doi.org/10.1016/j.neuron.2022.12.028 | PubMed

Lake BM, Salakhutdinov R, Tenenbaum JB (2015) Human-level concept learning through probabilistic program induction. *Science* **350**:1332-1338 https://doi.org/10.1126/science.aab3050 | PubMed

Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ (2017) Building machines that learn and think like people. *Behavioral and Brain Sciences* **40** https://doi.org/10.1017/S0140525X16001837 | PubMed

Lehnert L, Littman ML, Frank MJ (2020) Reward-predictive representations generalize across tasks in reinforcement learning. *PLoS Computational Biology* **16**:1-27 https://doi.org/10.1371/journal.pcbi.1008317 | PubMed

Luettgau L, Erdmann T, Veselic S, Stachenfeld KL, Kurth-Nelson Z, Moran R, Dolan RJ (2024) Decomposing dynamical subprocesses for compositional generalization. *Proceedings of the National Academy of Sciences* **121** https://doi.org/10.1073/pnas.2408134121 | PubMed

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

Luettgau L, Tempelmann C, Kaiser LF, Jocham G (2020) Decisions bias future choices by modifying hippocampal associative memories. *Nature Communications* **11**:3318 https://doi.org/10.1038/s41467-020-17192-7

Luyckx F, Nili H, Spitzer B, Summerfield C (2019) Neural structure mapping in human probabilistic reward learning. *eLife* **8**:e42816 https://doi.org/10.7554/eLife.42816 | PubMed

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

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

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

Murphy GL (1988) Comprehending Complex Concepts. *Cognitive Science* **12**:529-562 https://doi.org/10.1207/s15516709cog1204_2

Nieh EH, Schottdorf M, Freeman NW, Low RJ, Lewallen S, Koay SA, Pinto L, Gauthier JL, Brody CD, Tank DW (2021) Geometry of abstract learned knowledge in the hippocampus. *Nature* **595**:80-84 https://doi.org/10.1038/s41586-021-03652-7 | PubMed

Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N (2014) A toolbox for representational similarity analysis. *PLoS Computational Biology* **10**:e1003553 https://doi.org/10.1371/journal.pcbi.1003553 | PubMed

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

O'Donnell TJ, Goodman ND, Tenenbaum JB (2009) Fragment Grammars: Exploring Computation and Reuse in Language. www.csail.mit.edu

O'Donnell TJ, Snedeker J, Tenenbaum JB, Goodman ND (2011) Productivity and Reuse in Language. In: 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 | PubMed

Reber AS (1967) Implicit Learning of Artificial Grammars 1. *Journal of Verbal Learning and Verbal Behavior* **6** https://doi.org/10.1016/s0022-5371(67)80149-x

Redish AD (2016) Vicarious trial and error. *Nature Reviews Neuroscience* **17**:147-159 https://doi.org/10.1038/nrn.2015.30 | PubMed

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

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

Rubino V, Hamidi M, Dayan P, Wu CM (2023) Compositionality under time pressure. In: Proceedings of the Annual Meeting of the Cognitive Science Society. **45** https://doi.org/10.31234/osf.io/z2648

Schapiro AC, Rogers TT, Cordova NI, Turk-Browne NB, Botvinick MM (2013) Neural representations of events arise from temporal community structure. *Nature Neuroscience* **16**:486-492 https://doi.org/10.1038/nn.3331 | PubMed

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**:4885-4897.e14. https://doi.org/10.1016/j.cell.2023.09.004 | PubMed

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**:1214-1226.e8. https://doi.org/10.1016/j.neuron.2021.02.004 | PubMed

Shepard RN (1987) Toward a Universal Law of Generalization for Psychological Science. *Science* **237** https://doi.org/10.1126/science.3629243 | PubMed

Sherman BE, Graves KN, Turk-Browne NB (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 | PubMed

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.2200800119

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

Story GW, Smith R, Moutoussis M, Isabel M Berwian, Nolte T, Bilek E, Jenifer Z, Siegel JZ, Dolan RJ (2024) A Social Inference Model of Idealization and Devaluation. *Psychological Review* **131**:749-780 https://doi.org/10.1037/rev0000430 | PubMed

Teichmann L, Grootswagers T, Carlson T, Rich AN (2018) Decoding digits and dice with magnetoencephalography: evidence for a shared representation of magnitude. *Journal of Cognitive Neuroscience* **30**:999-1010 https://doi.org/10.1162/jocn_a_01257 | PubMed

Tsividis PA, Loula J, Burga J, Foss N, Campero A, Pouncy T, Gershman SJ, Tenenbaum JB (2021) Human-Level Reinforcement Learning through Theory-Based Modeling, Exploration, and Planning. *arXiv* https://doi.org/10.48550/arxiv.2107.12544

Turk-Browne NB, Isola PJ, Scholl BJ, Treat TA (2008) Multidimensional Visual Statistical Learning. *Journal of Experimental Psychology: Learning Memory and Cognition* **34**:399-407 https://doi.org/10.1037/0278-7393.34.2.399 | PubMed

Whittington JCR, McCaffary D, Bakermans JJW, Behrens TEJ (2022) How to build a cognitive map: insights from models of the hippocampal formation. *arXiv* https://doi.org/10.48550/arXiv.2202.01682

Whittington JCR, Muller TH, Mark S, Chen G, Barry C, Burgess N, Behrens TEJ (2020) The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell* **183**:1249-1263.e23. https://doi.org/10.1016/j.cell.2020.10.024 | PubMed

Yang GR, Joglekar MR, Song HF, Newsome WT, Wang XJ (2019) Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience* **22**:297-306 https://doi.org/10.1038/s41593-018-0310-2 | PubMed

Zhang M, Yu Q (2024) The representation of abstract goals in working memory is supported by task-congruent neural geometry. *PLoS Biology* **22**:e3002461 https://doi.org/10.1371/journal.pbio.3002461 | PubMed

Zheng XY, Hebart MN, Grill F, Dolan RJ, Doeller CF, Cools R, Garvert MM (2024) Parallel cognitive maps for multiple knowledge structures in the hippocampal formation. *Cerebral Cortex* **34** https://doi.org/10.1093/cercor/bhad485 | PubMed

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

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

Kern S, Nagel J, Gerchen MF, Guersoy C, Meyer-Lin-Denberg A, Kirsch P, Dolan RJ, Gais S, Feld GB (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. *Neuron* **111**:454-469 https://doi.org/10.1016/j.neuron.2022.12.028 | PubMed

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

Liu Y, Dolan RJ, Kurth-Nelson Z, Behrens TEJ (2019) Human Replay Spontaneously Reorganizes Experience. *Cell* **178**:1-13 https://doi.org/10.1016/j.cell.2019.06.012 | PubMed

Liu Y, Mattar MG, Behrens TEJ, Daw ND, Dolan RJ (2021) Experience replay is associated with efficient nonlocal learning. *Science* **372** https://doi.org/10.1126/science.abf1357 | PubMed

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

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**:4885-4897.e14. https://doi.org/10.1016/j.cell.2023.09.004 | PubMed

Peer reviews

Reviewer #1 (Public review):

Summary:

The paper presents an elegant task designed to investigate humans' ability to generalize knowledge of learned graph structures to new experiences that share the same structure but are built from different stimuli. Using behavior and MEG recordings, the authors test evidence for neural representation and application of structural knowledge.

Comments on revised version.

The authors have addressed my concerns and have largely toned-down the overstatement of their findings which was pervasive in the previous version of the manuscript. However, the revised title still seems rather misleading given the findings presented. I suggest the authors revise this to: 'A neural mechanism for transfer of abstract structural knowledge in humans'.

https://doi.org/10.7554/eLife.107162.2.sa3

Reviewer #2 (Public review):

The authors track structural role representations during stimulus processing, which is relevant to study how prior experience shapes new learning. They employ a sequence learning task based on graph factorization and record brain activity using MEG to investigate whether graph factors are reused during learning and inference in a new graph. MEG data were recorded during passive stimulus presentation, while behavior is assessed in a subset of probe trials testing either experienced or inferred successions in the graph. The authors

report neural similarity among stimuli conditional on their role in subprocesses, and a correlation between evidence for role representations and performance in experience probes, but not in inference probes.

The interpretation of "knowledge transfer" in the title and abstract is, however, constrained by the absence of a corresponding relationship with performance in the inference probes, which provide the behavioral measure of transfer. While the authors discuss why role representations might relate more closely to experience than inference probes in this task, this distinction makes the link between the neural findings and transfer weaker than the current framing suggests. (E.g., "was positively related to behavioral success in subprocess knowledge transfer" in the abstract could be misleading)

More broadly, the framing of the results as a mechanism for structure transfer or structural scaffolding appears stronger than warranted by the evidence. The findings indicate representations consistent with reusable roles, but do not establish that these representations functionally facilitate learning and transfer. The claims should align with this distinction between evidence for representational reuse and structural scaffolding.

Also, apart from abstract roles, simpler categorical grouping of different stimuli belonging to the same subprocess could produce increased similarity.

Relatedly, although the MEG-behavior (experience probes) correlation is significant for Component 6 but not Component 4, the difference between the two correlations is not significant. Interpreting the correlation as selective to the structurally relevant graph is thus difficult to statistically substantiate.

Overall, the study provides a well-designed paradigm for investigating compositional representations during structural generalization at high temporal resolution. However, the findings do not provide sufficient evidence to frame them as a mechanism for structure transfer.

https://doi.org/10.7554/eLife.107162.2.sa2

Reviewer #3 (Public review):

Summary:

This study investigates how task components can be learned and transferred across different task contexts. The authors designed two consecutive sequence learning tasks, in which complex image sequences were generated from the combination of two graph-based structural "building blocks". One of these components was shared between the prior and transfer task environments, allowing the authors to test compositional transfer. Behavioral analyses using generalized linear models (GLMs) assessed participants' sensitivity to the underlying structure. MEG data were recorded and analyzed using classifications and feature representational similarity analysis (RSA) to examine whether neural similarity increased for stimuli sharing the same relational structure. The paper aims to uncover the neural dynamics that support compositional transfer during learning.

Strengths and weaknesses:

I found the methods and task design of this paper difficult to follow, particularly the way stimuli were constructed and how the experimental sequences were generated from the graph structures. These aspects would be hard to replicate without some clarification. I appreciate the integration of behavioral and neuroimaging data. The overall approach, especially the use of compositional graph structures in sequence learning, is interesting and could be used and revised in further studies in compositionality and transfer learning. I appreciated the authors' careful interpretation of their findings in the discussion. However, I

would have liked a similar level of caution in the abstract, which currently overstates some claims.

Comments on revised version.

No further comments. I thank the authors for their diligence in addressing my comments.

https://doi.org/10.7554/eLife.107162.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***eLife Assessment:***
>
> *This study provides valuable insights into humans' ability to generalize knowledge of learned graph structures to new experiences that share the same structure but are built from different stimuli. However, the evidence for the authors' claims is incomplete, with the main claims of structural generalization and compositionality only partially supported by MEG and behavioral data. This study will be of interest to cognitive neuroscientists studying structure learning and generalization.*

We thank the editor and reviewers for a positive evaluation of our work and constructive suggestions to improve our manuscript.

The reviewers' comments motivated textual modifications to the manuscript, changes and additions to figures and conduct of further analyses.

We believe that the changes made in response to the reviewers' comments and suggestions provide for a substantially improved revised manuscript.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public review):***
>
> *Summary of the paper:*
>
> *The paper presents an elegant task designed to investigate humans' ability to generalize knowledge of learned graph structures to new experiences that share the same structure but are built from different stimuli. Using behavior and MEG recordings, the authors test evidence for neural representation and application of structural knowledge.*

We appreciate this overall positive evaluation of our work and suggestions to strengthen the empirical support for the claims made in the paper.

> *(1) While the task design is elegant, it isn't clear to me that the data support all the claims made in the paper. I have detailed my concerns below.*
>
> *The authors claim that their findings reveal "striking learning and generalization abilities based on factorization of complex experiences into underlying structural elements, parsing these into distinct sub processes derived from past experience, and forming a representation of the dynamical roles these features play within distinct subprocesses." And "neural dynamics that support compositional generalisation, consistent with a structural scaffolding mechanism that facilitates efficient adaption within new contexts".*
>
> *(a) First, terms used in these example quotes (but also throughout the paper) do not seem to be well supported by data or the task design. For example, terms such as 'compositional generalisation' and 'building blocks' have important relevance in other*

> *papers by (some of) the same authors (e.g., Schwartenbeck et al., 2023), but in the context of this experiment, what is 'composition'?*

We agree that our use of composition is less technical and formal than the nested and recursive forms of composition using "building blocks" reported in previous work (e.g., Schwartenbeck et al., 2023). In this experiment we study composition of smaller graphs via superimposition of their outputs. The superimposed "compound" outputs can be viewed as outputs of a single, more complex, graph. This is a special case of all possible ways that graphs can be composed from other graphs, and does not depend on a complex syntax for their combination (although more complex combination rules, other than multiplication are theoretically possible). Nevertheless, in accordance with common definitions of compositionality, this leads to a generative setup, given that a huge number of graphs can be expressed as the product of smaller graphs in this way.

Figure panel (Fig 1A) illustrates this precise point. Importantly, we argue this kind of composition is ubiquitous in real-world learning scenarios where multiple independent processes contribute to the holistic experience of a situation.

Based on the reviewer's suggestion we have now amended the manuscript introduction to render the distinction between different definitions, and usage, of the term composition more salient.

It is our understanding that this form of flexible, combinatorial, reuse of factorized prior knowledge is considered a specific instance of composition. Indeed, this understanding is supported by a cognitive-computational neuroscience literature that adheres to similar conceptualizations of compositionality, e.g., Barron et al., 2013 Nat Neuro; Rubino et al., 2023 PsyArxiv; Luettgau et al., 2024 PNAS.

We add a paragraph to the introduction section elaborating on what we mean by graph factorization and its relevance for questions of compositional generalization (reproduced below for convenience).

To address these questions, we developed a sequence learning paradigm based on a graph theoretical concept of graph factorization, emulating environments whose dynamics are the product of multiple simultaneous subprocesses. "Product graphs (Cartesian product), wherein 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 compelling and flexible test bed for probing compositional generalization of multiple (independent) and simultaneously unfolding subprocesses." Examples of such graph products and their factorizations, from a vast space of possible combinations, are provided in Figure 1A. Here we take inspiration from prior studies in neuroscience (e.g., Barron et al., 2013; Rubino et al., 2023; Schwartenbeck et al., 2023) and a prior cognitive 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).

> *(2) Can the authors demonstrate clear behavioural or neural evidence for compositional use of multiple graph structures, or alternatively remove reference to these terms? In the current paper, it seems to me that the authors are investigating abstract knowledge for singular graph structures (together with the influence of prior learning), as opposed to knowledge for the compound, more complex graph formed from the product of two simpler graphs.*

We thank the reviewer for this important clarification, and agree our language should more precisely distinguish what the task design requires from what our data directly demonstrate. The reviewer is correct that our analyses primarily provide evidence for the abstraction and transfer of a single previously-experienced graph structure to novel stimuli, rather than

direct evidence that participants combine two independently learned structures into a novel compound representation. We agree this distinction is important, and we have revised our language throughout the manuscript accordingly, reserving "compositional generalization" to describe the broader task structure (in which every compound stimulus is the joint product of two simultaneously evolving subprocesses, and the specific compound experienced during transfer learning was never observed during prior learning) and using more precise language: "decomposition," "abstraction," and "transfer" of single structural components to describe what our behavioral and neural evidence actually establishes. We have not claimed direct evidence that participants compositionally construct novel joint representations from two previously separate structures; rather, our evidence concerns the selective reuse of one such component when it recurs across learning phases, embedded in a novel pairing with another, entirely new structural component. We have adjusted the manuscript throughout and changed the title of the paper to “A neural mechanism for compositional structure transfer in humans”, to more adequately reflect the evidence presented.

> *(3) While I would like to be convinced that this data provides evidence for the transfer of abstract, structural knowledge, I think the authors either need to provide more convincing evidence or tone down their claims.*
>
> *Can the increase in neural similarity between stimuli mapping to the same abstract structural sub-process not be explained by temporal proximity in experiencing the transitions (e.g., Cai et al., 2016)? Indeed, behavior seems to be dominated by direct experience of the structure as opposed to applying abstract knowledge of equivalent structures (and, as a result, there is little difference in behavioural performance between experience and inference probes).*

The reviewer makes an astute observation while providing a constructive challenge. We agree that distinguishing transfer of abstract structural knowledge from effects driven by direct temporal experience is critical, and we appreciate this opportunity to clarify our reasoning on this point.

The reviewer suggests that increased neural similarity between stimuli that map to the same abstract structural sub-process might be explained by a temporal proximity in experience, as shown in prior work (e.g., Cai et al., 2016). We agree that temporal proximity is a potentially powerful determinant of representational similarity. Likewise, we do not dispute that temporal statistics of experience play an important role in shaping such neural representations. Nevertheless, we believe these accounts do not provide a sufficient explanation for our results, which we consider reflects a consequence of being exposed to the outputs produced by sequences drawn from an underlying graph structure and a prior learning of these sequences.

In our task, the temporal statistics of stimulus experience are fully dictated by the transition structure of the graph. Different graph structures necessarily impose different constraints on temporal proximity (e.g., how many steps it takes to revisit a given node, which nodes are encountered at similar temporal distances, etc.). Thus, if stimuli that occupy equivalent structural roles across different graphs come to be represented more similarly, this is precisely because learners have abstracted over the specific transition structure that governs these temporal relationships. In other words, consistent with our previous computational findings that models based on temporal difference learning mechanisms account well for behavioural response patterns (Luettgau et al., 2024 PNAS), temporal proximity is not a confound but instead is a likely mechanism through which structural knowledge is acquired. Critically, our key claims concern transfer across learning contexts, and not sensitivity to temporal proximity within a single experience stream. By design, the selective learning improvements observed in each group during transfer cannot be explained by temporal proximity alone, because the relevant temporal statistics in the transfer phase are novel and differ across groups. The fact that prior learning selectively facilitates subsequent learning

depending on an earlier experience of an abstract structure indicates participants are not simply tracking local temporal regularities, but instead are reusing previously learned structural knowledge to guide learning within a new context.

We agree that behavior in our task is strongly influenced by direct experience, and we do not claim that participants rely exclusively on abstract structural knowledge. Indeed, the relatively small differences between experience and inference probes are consistent with the view that abstraction complements and facilitates, rather than replaces, learning from direct experience. Our claims are therefore deliberately modest: we argue for the transfer of abstract structural knowledge that shapes learning dynamics and neural representations, not for behavior dominated by explicit inference over abstract graphs.

To avoid any criticism related to over-interpretation, we have toned down our language and clarify throughout introduction, results and discussion section of the manuscript that (i) temporal proximity is an intrinsic feature of graph structure, (ii) neural effects reflect (in part) abstraction over these structured temporal relationships, and (iii) that our results provide evidence for transfer of structural knowledge that biases learning, rather than for fully context-independent representations or strong behavioral generalization.

> *(4) The strongest evidence for neural representation of abstract task structures seems to be the increase in similarity by transition type. But this common code for 'transition type' is only observed for 6-bridge graphs and only for experienced transitions. There was no significant effect in inference probes. Therefore, there doesn't seem to be evidence for the application of a knowledge scaffold to facilitate transfer learning. Instead, the data reflects learning from direct experience and not generalisation.*

We appreciate the reviewer's point, and agree it deserves a careful investigation. We carefully considered how this correlation could be interpreted under a non-transfer, pure direct-experience account. Our interpretation is that even if all graph learning relies entirely on direct experience, a correlation is still expected due to individual differences in learning ability. Under this framework, higher-ability participants are simply more likely to directly discover the same structural component (the 6-bridge) across both sessions (prior and transfer learning), which in turn supports higher MEG transition decodability. Concurrently, these same participants will exhibit higher behavioral accuracy due to their superior learning capacity. Importantly, under this account, these higher-ability participants should also show higher behavioural accuracy during the transfer phase for the 4-cycle component, predicting (a) a positive correlation between our transition decoding signal and 4-cycle behavioural accuracy, and (b) a positive correlation between 4-cycle and 6-bridge behavioural accuracy themselves, within each prior learning group.

We tested both predictions directly. For (a), in the 6-bridge prior group, average decoding strength at timepoints showing a selective significant group difference correlated with experience probe performance on the 6-bridge graph during transfer learning ($\rho = .48$, $p = .016$, surviving correction for two comparisons, $\alpha = .025$), but the corresponding correlation with 4-cycle behavioural accuracy did not reach significance ($\rho = .38$, $p = .071$). We acknowledge, however, that a direct comparison of these two dependent correlations (Steiger's test) did not reach significance ($p = .66$).

For (b), we directly correlated 4-cycle and 6-bridge transfer accuracy within each prior learning group. For experience probes, this correlation was significant in the 4-cycle prior group ($\rho = .51$, $p = .010$) but not in the 6-bridge prior group ($\rho = .24$, $p = .257$); for inference probes, the pattern reversed, with a significant correlation in the 6-bridge prior group ($\rho = .42$, $p = .035$) but not the 4-cycle prior group ($\rho = .16$, $p = .452$). In neither probe type did a direct comparison of the two groups' correlation coefficients reach significance (experience probes: $Z = -1.06$, $p = .291$; inference probes: $Z = .95$, $p = .340$). Given that two of these four single-group correlations were nominally significant (without correction for multiple

comparisons) while the corresponding between-group comparisons were not, and given the inconsistency in which group/probe-type combination shows the effect, we consider this pattern inconclusive rather than supportive of a clear general-ability signal linking performance on the two graph components.

We therefore cannot statistically rule out the general-ability account with the current sample, and have revised our interpretation accordingly: rather than claiming strict specificity to the 6-bridge structure, we now describe the correlation as selective to 6-bridge experience probes in this sample, while acknowledging that distinguishing structural reuse from a more general learning ability to represent novel relational structure remains an open question. We have added the decoding-strength comparison as a new Supplementary Figure 5, showing decoding-strength correlations with both 4-state and 6-bridge graph performance side by side, and report the direct behavioural correlation between graph components in the text, and have revised the relevant main-text passage to match these updated values.

We also wish to address why this correlation was selective to experience probes and did not extend to inference probes. We agree that, under an assumption of perfectly learned and context-independent abstract representations, one might expect performance on experience and inference probes to relate equally to the decoding signal. We believe the reviewer's concern is valid only under an assumption of perfect abstraction. In our task, abstract roles correspond to latent structural variables that are learned gradually and with uncertainty. Under such conditions, role inference is inherently context-dependent: the role of a given stimulus is not inferred in isolation, but in relation to the compound within which it appears. For example, if one element of a compound is already known with high confidence to belong to a particular subgraph, the role of the second element can be inferred more reliably; in contrast, when both elements are themselves uncertain, role inference becomes less precise. Critically, learning is unlikely to be uniform across compounds, as structural knowledge may be acquired with variable precision for different stimulus pairings depending on their frequency, co-occurrence statistics, and position within a sequence. This variability in learning precision means we would expect corresponding variability in a participant's ability to infer dynamical roles across different compound types, and consequently in their ability to make accurate predictions about those compounds. Under this account, our MEG signal, which indexes role-based representational structure specifically for experienced compounds, most directly reflects knowledge precision for those items; a correlation with inference probe accuracy requires an additional inferential step from experienced to novel compounds, introducing noise proportional to uncertainty in the underlying learned representations. We have revised the Discussion to make this reasoning more explicit.

> *(5) The authors frequently suggest that they are providing insight into temporal dynamics, but there is no mention of particular oscillations or particular temporal sequences of neural representation that support task performance.*

We thank the reviewer for highlighting this. We agree our use of the term "temporal dynamics" is potentially misleading, and we have now revised the manuscript accordingly. Our analyses did set out to identify specific neural oscillations, rhythmic processes, or temporally ordered stages of computation that support task performance. Rather, time-resolved MEG analyses is a method that allows us to track when representational content is expressed during stimulus processing, serving three research questions about subprocess-specific representational similarity, dynamical role representations, and relationship to behavioral transfer performance; as opposed to characterizing the temporal organization of underlying computations per se.

We have now removed or replaced references to "temporal dynamics" and "neural dynamics" throughout the manuscript where these might be interpreted as claims about oscillatory or sequentially structured neural mechanisms. Instead, we now frame our conclusions in terms of representational content and its modulation by prior learning experience. We believe this

aligns our language with what the data actually support and removes potential confusion regarding the scope of our claims.

> *(6) Regardless of point (b), can the authors provide more convincing evidence for a graph structure being represented per se (regardless of whether this representation is directly experienced or inferred)? From Figure 3C, it seems that the model RDM doesn't account for relative distance within the graph. Do they see evidence for distance coding? Can they reconstruct the graph from representational patterns using MDS?*

We thank the reviewer for raising this point. The reviewer is correct that the model RDM shown in (previous) Figure 3C does not specifically capture relative distance within the graph and asks whether there is evidence for distance coding or whether the graph can be reconstructed using techniques such as MDS. We agree that distance-based representations are one possible signature of graph knowledge, with such effects reported in a prior literature. However, our theoretical focus is on an investigation of a more functionally relevant aspect of graph structure learning and reuse of abstracted knowledge: the representation of graph nodes according to their structural roles or positions within a graph.

To address the question more directly, we ran an additional sensor-level RSA analysis using the localizer task data, comparing representational changes from PRE to POST learning. Specifically, we examined whether stimuli in the 6-bridge graph that map onto the same structural roles (e.g., boundary, intermediate, central) within the relevant graph structures show greater neural similarity from before to after learning, and whether this effect depends on the graph structure experienced during prior learning. This new analysis revealed a cluster-based, permutation-corrected, interaction effect between prior learning group and graph factor, consistent with a representational similarity increase selectively for stimuli sharing the same structural role in a graph structure relevant to each group's prior experience. Importantly, this effect occurs at similar time points as our earlier RSA effects based on overall stimulus-set similarity, suggesting that both analyses capture related aspects of the same underlying representational processes.

We believe this role-based representational similarity provides more direct evidence for graph-structured representations and their transfer across learning than distance-based or MDS reconstructions in our task context. First, distance within the graph is necessarily correlated with temporal proximity of experience, making it difficult to disentangle distance coding from local transition statistics. Second, multiple distinct graph structures used in our experiment (e.g., cycles, lines, and bridges) share overlapping distance relationships but differ in their higher-order structural roles. Our analysis therefore targets those aspects of structure that are invariant across surface features and that are most relevant for transfer across learning contexts.

For these reasons, while we do not claim we can reconstruct full graph geometry from neural patterns, our results are consistent with an organization of neural representations according to abstract structural positions within a graph, that is independent of specific stimulus identities. We have added a new panel to Figure 3 illustrating this role-based RSA effect and also expanded the Methods and Results sections to describe this analysis in detail. We believe this addition provides more convincing evidence that participants represent graph structure at an abstract level, even when that structure is not explicitly experienced in the same form during the localiser task. We thank the reviewer for encouraging this more specific analysis, as it significantly strengthens our claims.

"To test whether localiser representations of abstract graph structure generalised across sets experienced in the main task, we performed RSA separately for PRE and POST localisers. Structural roles were derived from each participant's learning-phase transition sequences. For 4-state graphs, nodes were assigned to two matched role bins (boundary-like vs central-like) based on first-encounter order within each phase. For 6-state graphs, roles were

determined either from node degree (6-line: boundary, intermediate, central). For 6-bridge graphs we used a predefined structural template. Specifically, the first eight valid transitions in the learning phase were mapped onto a fixed sequence of structural roles (central–intermediate–boundary–central–central–intermediate–boundary–central). Each node was assigned a role based on the majority role it occupied across its occurrences in these eight steps (ties resolved in favour of central > intermediate > boundary). For the localiser data, at each time point and sensor we estimated condition-specific activations using a GLM with 20 dummy-coded regressors (10 identities × 2 sets). This yielded sensor-wise beta patterns for each condition. Cross-set representational dissimilarities were computed using an L1 distance metric (mean absolute difference across sensors), resulting in 4×4 and 6×6 cross-set RDMs at each time point. We then summarised each RDM according to structural role: within-role distances (same role across sets), between-role distances (different roles), and their difference (delta = between − within). To control for unequal pair counts, between-role comparisons were subsampled to match the number of within-role pairs (200 repetitions). This produced time-resolved structural generalisation indices for 4-state and 6-state graphs, which were submitted to group-level statistical analyses as described above."

> *(7) In general, the figures are not very clear, and the outcome from statistical tests is not graphically shown. The paper would be easier to digest if, for example, Figures 1-2 were made clearer and statistical significance relative to chance were indicated throughout. To give two examples: (i) Figure 1 should clearly indicate what is meant by observed and held-out transitions and whether it is just the transition or also the compound that is new to the participant. (ii) Figure 2D-E could be shown with relevant comparisons and simpler statistical comparisons. Currently, it is hard to follow without carefully reading the legend.*

We have taken the reviewer's point on board and have revised the figures to improve clarity. Specifically, we have improved labeling and visual annotation in Figure 1 to clearly distinguish observed versus held-out transitions and to highlight the key experimental manipulations. We have also simplified the presentation of behavioral results in Figure 2D–E and made the most relevant statistical comparisons more explicit, with statistical outcomes now indicated directly in the figures where appropriate.

We believe these changes improve readability and reduce the burden on the reader, while still relying on the figure captions to convey essential methodological details and interpretation, following standard practice. See Figure 1 and Figure 2.

> ***Reviewer #1 (Recommendations for the authors):***
>
> *(1) For the classifier: what was the overall accuracy of the classifier when tested on held-out examples of the same stimuli? Relative to chance? Relative to entirely different stimuli that have the same transition type (e.g., line 424)?*
>
> *(2) In the Results paragraph 2 can you state how many subjects had prior exposure to 4-state vs how many had prior exposure to 6-state?*

This has now been added these data to the results section, reproduced below for convenience:

In brief, participants "(N = 50, N = 25 subjects per condition)" performed a sequence learning task, observing image sequences produced by walks on product graphs constructed from smaller graph factors (Fig. 1A and B).

> ***Reviewer #2 (Public review):***
>
> *Summary:*

> *The authors aimed to investigate the temporal dynamics of how prior experiences shape learning in new complex environments by examining whether the brain reuses abstract structural components from those experiences. They employed a sequence learning task based on graph factorization and recorded neural activity using magnetoencephalography (MEG) to investigate how the underlying graph factors are reused to support learning and inference in a new graph. MEG data was derived from passive stimulus presentation trials, and behavior was assessed through a small number of probe trials testing either experienced or inferred successions in the graph. Representational similarity analysis of the MEG data was performed at a quite aggregated level (the principal components explaining 80% of the variance). The authors report (1) enhanced neural similarity among stimuli that belong to the same graph-factor as well as (2) a correlation between abstract role representations, corresponding to particular positions in the graph, and performance in experience-probes but not in inference-probes.*
>
> *Overall, the findings are congruent with prior research on neural correlates of structural abstraction. They offer an elegant, well-suited task design to study compositional representations, replicating the authors' earlier finding and providing temporal information on structural generalisation in a sequence learning task.*

We thank the reviewer for this positive evaluation of our work and insightful comments and suggestions.

> *(1) The first finding is considered evidence for representational alignment of the graph factors. However, alignment seems to be just one possible arrangement underlying the increased similarity between stimuli of the same vs different graph factors. For instance, a simple categorical grouping of stimuli belonging to the same graph, rather than their structural alignment, could also underlie the reported effect. The wording should be adjusted to avoid over interpretation.*

We agree with the reviewer that increased neural similarity between stimuli associated with the same graph factor does not uniquely imply representational alignment in its strongest sense, and could in principle also reflect a more general categorical grouping of stimuli based solely on belonging to the same graph. We revised the manuscript to tone down the wording and avoid overinterpretation, replacing references to “representational alignment” with more neutral language that refers to increased within-factor representational similarity.

At the same time, we believe that additional analyses introduced in the revised manuscript constrain alternative accounts. In particular, our new analysis on localiser data, examining representational changes from PRE to POST learning for stimuli that map onto the same structural roles within relevant sub-process graph structures. This analysis showed the presence of a cluster-based, permutation-corrected, interaction effect between group and graph factor, indicating that representational similarity increases selectively for stimuli sharing the same structural role, rather than simply increased similarity for all stimuli associated with the same graph.

Because these role-based groupings cut across specific stimulus identities and cannot be reduced to a single categorical label for an entire graph, we believe this result provides specific evidence that neural representations are shaped by internal structural properties of the graph, rather than reflecting the emergence of a coarse categorical distinction between graph types. We have added this analysis as a new panel to Figure 3 and expanded the Results and Methods sections accordingly (reproduced below for convenience):

“The observed increase in neural similarity between stimuli associated with the same graph factor could, in principle, reflect a more general categorical grouping of stimuli belonging to

the same graph factor, rather than reflecting an abstract graph representation. To address this alternative account, we examined whether neural similarity in the localizer task is specifically structured according to the graph-theoretic role each stimulus occupies within its graph, using a finer-grained theoretical RDM that distinguishes between boundary, intermediate, and central positions. Structural roles were derived from each participant's learning task sequences and defined by graph topology. For Component 6, roles were assigned using a bridge-graph traversal template (for Condition 6, whose prior learning involved the bridge graph) or degree-based assignment (for Condition 4, whose prior learning involved a line graph): boundary nodes were defined as nodes with degree 1 (graph periphery); intermediate nodes as neighbours of boundary nodes; and central nodes as the remaining, more densely connected nodes. For Component 4, two roles were distinguished, boundary-like nodes (those encountered first and last in the sequence traversal) and central-like nodes (those encountered second and third), reflecting structural position within the graph rather than arbitrary temporal sequence. For further details of role assignment, see Methods. This analysis revealed a cluster-based permutation-corrected interaction (P_{FEW} < 0.05, non-parametric permutation test, Figure 3E) between prior learning group and graph factor at overlapping time windows with the previously reported PCA and sensor-level analyses investigating feature-mappings to abstract graph factors. We did not observe significant and multiple comparisons corrected prior learning group and graph factor main effects (all P_{FEW} > 0.90, non-parametric permutation test). These results indicate that representational similarity increased selectively for stimuli sharing the same structural role in the graph structure relevant to each group's prior experience. Critically, this effect was not observed uniformly for all stimuli associated with the same graph, arguing against a purely categorical grouping explanation. Instead, these findings provide evidence that learning induces representations that reflect higher-order relational structure, capturing graph-theoretic roles that are invariant across surface features of the embedding stimuli and support transfer across learning contexts.

Taken together, our findings indicate a task-induced, learning-related increase in neural similarity for stimuli sharing a structural role in subprocesses common across task phases and is compatible with a structural scaffolding mechanism, though we do not claim these representations are fully context-independent.

"To test whether localizer representations of abstract graph structure generalised across sets experienced in the main task, we performed RSA separately for PRE and POST localisers. Structural roles were derived from each participant's learning-phase transition sequences. For 4-state graphs, nodes were assigned to two matched role bins (boundary-like vs central-like) based on first-encounter order within each phase. For 6-state graphs, roles were determined either from node degree (6-line: boundary, intermediate, central). For 6-bridge graphs we used a predefined structural template. Specifically, the first eight valid transitions in the learning phase were mapped onto a fixed sequence of structural roles (central–intermediate–boundary–central–central–intermediate–boundary–central). Each node was assigned a role based on the majority role it occupied across its occurrences in these eight steps (ties resolved in favour of central > intermediate > boundary). For the localiser data, at each time point and sensor we estimated condition-specific activations using a GLM with 20 dummy-coded regressors (10 identities × 2 sets). This yielded sensor-wise beta patterns for each condition. Cross-set representational dissimilarities were computed using an L1 distance metric (mean absolute difference across sensors), resulting in 4×4 and 6×6 cross-set RDMs at each time point. We then summarised each RDM according to structural role: within-role distances (same role across sets), between-role distances (different roles), and their difference (delta = between − within). To control for unequal pair counts, between-role comparisons were subsampled to match the number of within-role pairs (200 repetitions). This produced time-resolved structural generalisation indices for 4-state and 6-state graphs, which were submitted to group-level statistical analyses as described above."

> *(2) The second finding of abstract role representations is indeed expected for structural generalisation. While the data presents an interesting indication, its interpretability is constrained by a lack of testing for generalization of the effect to other graph structures (e.g., to rule out graph-specific strategies) as well as the absence of a link to transfer performance in inference-probes. The authors argue that the experienced transitions the classifier was trained on might be more similar in process to the experience-probes than the inference-probes. However, as inference-probes are the key measure of transfer, one could argue that if abstract role representations truly underlie transfer learning, they should be evident in the common neural signal.*

The reviewer provides a thoughtful comment and we agree that, under an assumption of perfectly learned and context-independent abstract representations, one might expect this would be related equally to performance in experience and inference probes. Below, we clarify the scope of our claims, including presenting additional neural evidence for abstract role representations, and argue that an absence of a strong neural-behavioral correlation for inference probes is not inconsistent with our proposed abstract structural knowledge account of transfer learning.

First, we acknowledge that our experiment tests generalization only across two graph families, rather than across a broader set of graph structures. While we think it unlikely that our findings are idiosyncratic to cycles or bridge-like graphs, a we now limit our claims and emphasise a future need for testing generalization across additional graph families (see discussion section as reproduced below)

“We concede that the present experiment examines generalization across two families of graph structures, rather than across a broader space of possible graphs. While we consider it unlikely that our findings are idiosyncratic to cycles, lines and bridge-like graphs, we acknowledge that further work will be needed to support abstraction beyond the limited graph structures studied here.”

Second, to provide more direct neural evidence that abstract structural roles within the studied graphs are represented per se, we add a new analysis that uses independent localizer task data. In this analysis, we performed a sensor-level RSA contrasting pre- and post-learning representations. Specifically, we test whether stimuli that occupied the same role within learned graph substructures became more neurally similar after learning. This localizer RSA revealed a cluster-corrected interaction effect between prior learning group and graph factor, expressed in an increased representational similarity for stimuli mapping onto equivalent structural roles, independent of specific stimulus identities and outside the task context of the main learning phase. This result supports the proposal that participants form abstract, role-based, representations of graph structure where these representations are not tied to a particular transition sequence or probe type. This finding adds an important insight in that it goes beyond establishing role representations for the 6-bridge graph factor, enabling us to make claims about generalization of structural knowledge of the 4-cycle graph factor in the 4-cycle prior condition (thereby to some extent showing a wider generalization across more structural forms).

“To test whether localizer representations of abstract graph structure generalised across sets experienced in the main task, we performed RSA separately for PRE and POST localisers. Structural roles were derived from each participant’s learning-phase transition sequences. For 4-state graphs, nodes were assigned to two matched role bins (boundary-like vs central-like) based on first-encounter order within each phase. For 6-state graphs, roles were determined either from node degree (6-line: boundary, intermediate, central). For 6-bridge graphs we used a predefined structural template. Specifically, the first eight valid transitions in the learning phase were mapped onto a fixed sequence of structural roles (central–intermediate–boundary–central–central–intermediate–boundary–central). Each node was

assigned a role based on the majority role it occupied across its occurrences in these eight steps (ties resolved in favour of central > intermediate > boundary). For the localiser data, at each time point and sensor we estimated condition-specific activations using a GLM with 20 dummy-coded regressors (10 identities × 2 sets). This yielded sensor-wise beta patterns for each condition. Cross-set representational dissimilarities were computed using an L1 distance metric (mean absolute difference across sensors), resulting in 4×4 and 6×6 cross-set RDMs at each time point. We then summarised each RDM according to structural role: within-role distances (same role across sets), between-role distances (different roles), and their difference (delta = between − within). To control for unequal pair counts, between-role comparisons were subsampled to match the number of within-role pairs (200 repetitions). This produced time-resolved structural generalisation indices for 4-state and 6-state graphs, which were submitted to group-level statistical analyses as described above."

Third, in relation to the relationship between neural role representations and transfer performance during inference probes, we believe the reviewer's concern is valid only under an assumption of perfect abstraction. In our task, abstract roles correspond to latent structural variables that are learned gradually, and with uncertainty. Under such conditions, role inference is inherently context-dependent: the role of a given stimulus is not inferred in isolation, but in relation to the compound within which it appears. For example, if one element of a compound is already known with high confidence to belong to a particular subgraph, the role of the second element can be inferred more reliably. In contrast, when both elements are themselves uncertain, role inference becomes less precise. Critically, learning is unlikely to be uniform across compounds, as structural knowledge may be acquired with variable precision for different stimulus pairings depending on their frequency, co-occurrence statistics, and position within a sequence. This variability in learning precision means we would expect corresponding variability in a participant's ability to infer dynamical roles across different compound types, and consequently in their ability to make accurate predictions about those compounds.

Our MEG signal indexes role-based representational structure specifically for experienced compounds, and most directly reflects the precision of structural knowledge for those items. Therefore a correlation between a MEG signal and accuracy on inference probes is expected to be weaker, since it requires an additional inferential step from experienced to novel compounds, a step that is likely to introduce additional noise proportional to the uncertainty in the learned representations. Ideally we would expect such a correlation to be present for inference probes as well, since better structural knowledge about experienced compounds should in principle support better generalisation to novel ones. However, we consider we are insufficiently powered to detect this weaker relationship and interpret an absence of a significant correlation for inference probes as not inconsistent with a structural scaffolding account.

Finally, we emphasize that the absence of a strong correlation between aggregate neural decoding measures and aggregate inference-probe performance should not be taken as evidence against abstract learning. Inference performance is likely to be supported by multiple sources of information (including partial structural knowledge and heuristic strategies), whereas neural decoding captures the stability of representational geometry at a population level. The newly added localizer RSA results demonstrate that abstract role representations are present in the neural signal, while at the same time their task-dependent expression may not always be directly mirrored at the level of inference-probe behavior.

We have an added a paragraph caveating our findings to the discussion section (reproduced below for convenience):

"Addressing our third research question, whether the strength of identified neural representations relate to behavioral transfer, we found a positive relationship between decoding strength and experience probe performance that was numerically larger for the

structurally relevant (6-bridge) graph than for the irrelevant (4-cycle) graph, though this difference did not reach significance; no corresponding relationship was evident for inference probes." Thus, a behavioral accuracy group difference in predicting successive states for the 6-bridge inference probe performance was of negligible magnitude and did not align with the robust effect observed for experience probes (Fig. 2D), where the latter replicates and extends a previous finding using a 6-state cycle graph and predicted by computational simulations using successor feature models (Luettgau et al., 2024). We speculate that the disparity between probe types is likely due to experience probes and inference probes engaging dissociable cognitive processes – the former relying on predictive representations acquired from experience, the latter relying on an ability to perform mental operations on learned representations of independent feature transitions. This may increase the likelihood of errors due to complexities of the inferred state spaces. However, the observed differences could also relate to previously reported effects of explicit knowledge on performance levels (Luettgau et al., 2024) or be driven by capacity limitations in working memory taxing larger subprocesses more strongly (Collins et al., 2017; Collins & Frank, 2012). "It is also likely that structural knowledge is not acquired uniformly across compounds. Because abstract roles are learned gradually and with uncertainty, role inference is inherently context-dependent: the role of a given stimulus is not inferred in isolation, but in relation to the compound within which it appears. When one element of a compound is already known with high confidence, the role of the second element can be inferred more reliably; when both elements are themselves uncertain, inference becomes less precise. Learning precision may therefore vary across stimulus pairings depending on their frequency, co-occurrence statistics, and position within a sequence, and we would expect corresponding variability in participants' ability to infer dynamical roles, and hence to predict outcomes, across different compound types. This means that an MEG signal, which indexes role-based representational structure specifically for experienced compounds, most directly reflects knowledge precision for those items. A correlation with inference-probe accuracy requires an additional inferential step, from experienced to novel compounds, introducing noise proportional to the uncertainty in the learned representations. In principle, better structural knowledge about experienced compounds should support better generalisation to novel ones, but we might be insufficiently powered to detect this weaker relationship. We therefore interpret an absence of a significant correlation for inference probes not as inconsistent with a structural scaffolding account, and suggest that a null result here should be interpreted with caution rather than as evidence against it. More broadly, an absence of a strong correlation between aggregate neural decoding measures and aggregate inference-probe performance should not itself be taken as evidence against abstract learning. Inference performance is likely supported by multiple sources of information, including partial structural knowledge and heuristic strategies, whereas neural decoding captures the stability of representational geometry at the population level. The localizer RSA results reported above demonstrate that abstract role representations are present in the neural signal, even though their task-dependent expression may not always be directly mirrored in inference-probe behavior."

> *(3) The authors write, "we observed a qualitative pattern indicative of increased neural similarity between stimuli that adhered to the same underlying subprocess across task phases. (...) There was a statistically significant interaction effect of condition x graph factor spanning approximately 300 - 680 ms post-stimulus onset". I conclude there was no significant main effect of graph factor, but the relevant statistics are not reported. The authors should report and discuss the complete statistics.*

Thank you for this comment. The reviewer is correct in their conclusion that the main effects were not significant and we agree the main effect of graph factor and group (as well as their significance) should be reported. We now do so in the figure caption of Figure 3D, as well as in the results section. The section reads as follows:

We did not observe prior learning group and graph factor main effects (all PFWE > 0.118, non-parametric permutation test).

> *(4) The RSA is performed on highly aggregated data (the PCs that explained 80% of the variance). Could the authors include their rationale for this choice (e.g. over-analysis of sensor-level data)? In case sensor-level analyses have been conducted as well, maybe there are comparisons or implications of the chosen approach that are useful to mention in the discussion. The authors should provide the average and distribution of the number of PCs underlying their analyses.*

We thank the reviewer for raising this important methodological concern. We agree that the use of PCA-based dimensionality reduction in RSA requires further justification. Our primary motivation for using PCA, and retaining components that jointly explain 80% of the variance, was based on denoising rather than aggregation per se. Sensor-level MEG data are high-dimensional and highly correlated, and representational analyses that operate directly on the full sensor space may be susceptible to noise amplification and overfitting, particularly when estimating distance measures across time. Retaining components up to a fixed variance threshold is a widely used approach in MEG decoding and RSA to preserve the dominant signal structure while attenuating sensor-specific noise. Importantly, this procedure discards only the lowest-variance 20% of the signal and on this basis does not impose strong constraints on the representational geometry.

To address the reviewer's concern more directly, and to ensure our findings are not driven by the PCA procedure itself, we conducted additional robustness checks. Specifically, we performed sensor-level RSA analyses without PCA, using multiple distance metrics (including L1 and L2 distances). These sensor-level analyses yielded highly similar temporal profiles and statistical effects to the PCA-based results, including for the key contrasts reported in the main text. We now include these results in a new supplementary figure and reference them explicitly in the Results section.

In the revised manuscript, we clarify that PCA was used as a principled denoising step, not as a means of imposing structure on the data, and we explicitly show that our conclusions are robust across analytic choices.

"Importantly, to verify that the chosen PCA approach did not impose artificial structure within the data, or introduce other potential confounds, we also conducted the above analyses on sensor level MEG data. Highly similar results were obtained, with consistent results for different similarity metrics (absolute distance (L1 norm) and Euclidean distance (L2 norm), Supplementary Figure 3)), demonstrating that our results and conclusions are robust across analytic choices."

Methods

"In addition to the PCA-based feature analysis, we performed a sensor-level RSA. At each time point, we estimated condition-specific activation patterns using a GLM applied separately to each sensor (dummy-coded stimulus identity). RDMs were then computed across stimuli using mean absolute difference (L1 norm) as the primary distance metric. For completeness, we additionally evaluated Euclidean distances (L2 norm), which yielded qualitatively similar results (Supplementary Figure 3)."

> *(5) While the paper is well-written overall, it would benefit from more explicitly identifying the concrete research question and advancing through the results. The authors state their aim as understanding the "temporal dynamics of compositional generalisation", revealing "at which moment during neural information processing are they assembled". They conclude with "providing evidence for temporally resolved neural dynamics that support compositional generalization" and "we show the neural dynamics (...) presented*

> *across different task phases...". It remains somewhat vague what specific insight about the process is provided through the temporal resolution (e.g., is the time window itself meaningful, if so, it should be contextualized; is the temporal resolution critical to dissociate subprocesses). The different task phases -initial learning and transfer- are the necessary conditions to investigate transfer learning, but do not by themselves offer a particularly resolved depiction of the process.*

We thank the reviewer for these thoughtful and constructive comments. We agree that in the original version of the manuscript our framing did not sufficiently specify the concrete insight that might be gained from neural data, nor did it clearly delimit the scope of our claims. Our primary research question did not relate to characterizing fine-grained temporal mechanisms or sequential subprocesses underlying compositional generalization (e.g., oscillatory dynamics, replay, or temporally ordered stages of computation). Rather, our goal is to determine whether abstract compositional structural representations are detectable neurally during stimulus processing, and how these representations are shaped by prior learning experience.

We address three specific research questions: whether subprocess-specific representational similarity emerges following learning that is specific to a subprocess experienced during prior learning; whether fine-grained dynamical role representations are acquired during stimulus processing; and whether their strength relates to a behavioral transfer success. Importantly, the time-resolved nature of the analyses is a feature of the method that allows us to track the expression of representational content at a sub-trial timescale, rather than an end in itself. We agree with the reviewer that the temporal resolution does not, in itself, resolve distinct computational subprocesses, nor does it provide a mechanistic decomposition of compositional operations. Accordingly, we have revised the manuscript to more explicitly articulate these three concrete research questions in the Introduction, as well as clarify that the reported time windows should be interpreted as markers of representational availability rather than discrete processing stages. We also tone down any language that might be interpreted as making claims about temporally structured mechanisms. Finally, we agree that the distinction between task phases (prior learning vs. transfer) is not intended to depict a detailed temporal process, but rather to provide the necessary contrast to test whether neural representations reflect abstract structural knowledge that goes beyond direct experience. We have revised the Results and Discussion to make this logic more explicit and to better guide the reader through how each analysis contributes to addressing the three research questions.

We believe these revisions better align our claims with the level of insight the data support and make the contribution of each analysis more legible to the reader.

> ***Reviewer #2 (Recommendations for the authors):***
>
> *(1) Could the authors explain in the manuscript how the held-out transitions were chosen?*

We agree that the stimulus construction and sequence generation logic were not described with sufficient clarity. In the revised manuscript we add an explicit description of how compound images were constructed from two independently varying visual features, each corresponding to one graph factor, and introduced the distinction between directly experienced, implied-but-unobserved, and never-valid transitions earlier in the text, so that the probe logic is easier to follow from the outset.

We reproduce parts of the revised methods section describing the held-out stimulus set below:

“[...] Because sequences were drawn from a subset of the product graph, only a portion of all 24 possible compound images (and their transitions) were encountered directly. Transitions

between compound images could thus be: (i) directly experienced in a sequence, (ii) never presented in a sequence but implied by the factorization structure, or (iii) transitions between compound images from different subgraphs, which were never valid. The held-out set of transitions described in (ii) and (iii) between compound stimuli (partially observed during sequence presentations, partially never observed during sequence presentations) provided a context for assessing knowledge of graph factorization."

> *(2) It might be easier to follow the results section (without going back to Figure 1) if the mentioning of specific graph factors (6-bridge, 4 cycle) would occasionally be accompanied by their function/condition in the experiment.*

Thanks for this suggestion, we have implemented this in the manuscript.

> ***Reviewer #3 (Public review):***
>
> *Summary:*
>
> *This study investigates how task components can be learned and transferred across different task contexts. The authors designed two consecutive sequence learning tasks, in which complex image sequences were generated from the combination of two graph-based structural "building blocks". One of these components was shared between the prior and transfer task environments, allowing the authors to test compositional transfer. Behavioral analyses using generalized linear models (GLMs) assessed participants' sensitivity to the underlying structure. MEG data were recorded and analyzed using classifications and feature representational similarity analysis (RSA) to examine whether neural similarity increased for stimuli sharing the same relational structure. The paper aims to uncover the neural dynamics that support compositional transfer during learning.*
>
> *I found the methods and task design of this paper difficult to follow, particularly the way stimuli were constructed and how the experimental sequences were generated from the graph structures. These aspects would be hard to replicate without some clarification. I appreciate the integration of behavioral and neuroimaging data. The overall approach, especially the use of compositional graph structures in sequence learning, is interesting and could be used and revised in further studies in compositionality and transfer learning. I appreciated the authors' careful interpretation of their findings in the discussion. However, I would have liked a similar level of caution in the abstract, which currently overstates some claims.*

We thank the reviewer for a positive evaluation coupled with helpful suggestions to improve the clarity of our presentation. We are particularly pleased the reviewer found the integration of behavioral and neuroimaging data, the compositional graph structure approach, and our interpretation of findings in the discussion to be particular strengths of the manuscript.

We address the two specific concerns raised below.

We agree that the stimulus construction and sequence generation logic were not described with sufficient clarity. In the revised manuscript we have: (i) added an explicit description of how compound images were constructed from two independently varying visual features, each corresponding to a single graph factor; (ii) defined the Cartesian product graph operationally, clarifying also that transitions between compound images are valid when one factor changes state while the other remains fixed; (iii) included a worked example tracing a sample sequence through both factors; and (iv) introduced the distinction between directly experienced, implied-but-unobserved, and never-valid transitions earlier in the text, so that the probe logic is easier to follow from the outset.

We reproduce parts of the revised methods section describing stimulus construction below:

“Compound images were constructed by overlaying two independently varying visual features drawn from separate feature sets. Each feature set corresponded to a single graph factor, such that the state of each graph factor independently determines one feature of the compound image. For example, a compound image in state (A, 3) would display feature A from the cyclic factor and feature 3 from the path factor, superimposed or juxtaposed into a single image.

The set of all compound images and their transitions was defined by the Cartesian product of two graph factors: transitions between compound images were valid if, and only if, one factor alone changed state according to its own transition rules while the other factor remained fixed.

Observed sequences of image compositions composed from 10 individual features. Each sequence consisted of 8 compound images with 12 compound images in total across sequences.

[...] For instance, in condition 1, a valid sequence might proceed: (A,1) → (A,2) → (A,3) → (B,3) → (C,3) → (C,4) → (D,4) → (D,5), where each step reflects a transition in either the cyclic (letter) or path (number) factor, but never both simultaneously.

[...] Because sequences were drawn from a subset of the product graph, only a portion of all 24 possible compound images (and their transitions) were encountered directly. Transitions between compound images could thus be: (i) directly experienced in a sequence, (ii) never presented in a sequence but implied by the factorization structure, or (iii) transitions between compound images from different subgraphs, which were never valid. The held-out set of transitions described in (ii) and (iii) between compound stimuli (partially observed during sequence presentations, partially never observed during sequence presentations) provided a context for assessing knowledge of graph factorization.”

We thank the reviewer for flagging the mismatch between the strength of interpretations made in the abstract and discussion section. We have revised the abstract to bring the language in line with the level of caution exercised in the discussion, avoiding overstated interpretations of the findings.

The revised abstract is reproduced below:

“A human ability to adapt to the dynamics of novel environments relies on abstracting and generalizing from past experiences. Previous research has focused on how humans generalize from isolated sequential processes, yet we know little about mechanisms that enable adaptation to more complex dynamics, including those that govern much everyday experience. Here, using a novel sequence learning task based on graph factorization, coupled with simultaneous magnetoencephalography (MEG) recordings, we ask whether reuse of experiential "building blocks" enable inference and generalization. Behavioral results were consistent with participants decomposing task experience into sub processes, abstracting their dynamical structure away from their sensory specifics and transferring these to a new task environment. Neurally, this transfer was associated with a learning-induced, condition-specific, increase in neural similarity among stimuli conditional on their structural role in abstract sub processes that were shared across task phases. Consistent with this, a complementary dynamical role decoding analysis revealed enhanced neural decodability for stimuli sharing identical structural transition types across experiential contexts, specifically when prior experience included the relevant graph structure. Decoding strength for these role representations was positively related to behavioral success in sub process knowledge transfer, a relationship that will require future independent replication. These findings

provide neural evidence consistent with the idea that a structural scaffolding mechanism supports generalization of experience to new contexts."

> *(1) While the introduction mentions brain areas implicated in the low-dimensional representation of task knowledge, the current study uses M/EEG and does not include source reconstruction. As a result, the focus is primarily on the temporal dynamics of the signal rather than its spatial origins. Although I am not suggesting that the authors should perform source reconstruction in this study, it would strengthen the paper to introduce the broader M/EEG literature on task-relevant representations and transfer. The same applies to behavioral studies looking at structural similarities and transfer learning. I encourage the authors to integrate relevant literature to better contextualize their results.*
>
> *Duan, Y., Zhan, J., Gross, J., Ince, R. A. & Schyns, P. G. Pre-frontal cortex guides dimension-reducing transformations in the occipito-ventral pathway for categorization behaviors. Current Biology 34, 3392-3404 (2024).*
>
> *Luyckx, F., Nili, H., Spitzer, B. & Summerfield, C. Neural structure mapping in human probabilistic reward learning. eLife 8, e42816 (2019). (This is in the references but not in the text).*
>
> *Zhang, M. & Yu, Q. The representation of abstract goals in working memory is supported by task-congruent neural geometry. PLoS biology 22, e3002461 (2024).*
>
> *L. Teichmann, T. Grootswagers, T. Carlson, A.N. Rich Decoding digits and dice with magnetoencephalography: evidence for a shared representation of magnitude Journal of cognitive neuroscience, 30 (7) (2018), pp. 999-1010*
>
> *Garner, K., Lynch, C. R. & Dux, P. E. Transfer of training benefits requires rules we cannot see (or hear). Journal of Experimental Psychology: Human Perception and Performance 42, 1148 (2016).*
>
> *Holton, E., Braun, L., Thompson, J., Grohn, J. & Summerfield, C. Humans and neural networks show similar patterns of transfer and interference during continual learning (2025).*

We thank the reviewer for providing these helpful suggestions. We have taken these references on board and now introduce these findings in the introduction of the manuscript (reproduced below for convenience):

"In parallel, M/EEG studies have shown that abstract, format-independent representations can emerge dynamically over time. For example, multivariate analyses of MEG data demonstrate cross-decoding of numerical magnitude across distinct symbolic formats (e.g., digits and dice), indicating a shared, format-invariant representation that unfolds with distinct temporal dynamics (Teichmann et al., 2018). More broadly, M/EEG evidence suggests that task-relevant information can be dynamically prioritized, compressed, and reformatted, indicative of a rapid temporal evolution of abstract and behaviorally relevant representations even when spatial localization remains coarse. Recent MEG data suggests that high-dimensional sensory inputs are progressively transformed into lower-dimensional, task-relevant, feature manifolds across distinct processing stages, reflecting dynamic gating and reformatting of information in support of flexible categorization (Duan et al., 2024). Complementary EEG–fMRI evidence also indicate abstract task goals are encoded in a low-dimensional neural geometry that emerges in frontal cortex and communicated to posterior regions to guide implementation of stimulus-specific representations, with the fidelity of this geometry predicting behavioral performance (Zhang & Yu, 2024). Likewise, converging behavioral and computational data indicates successful generalization depends on detecting and encoding abstract relational structure rather than stimulus surface features. In effect,

training enables transfer when task phases share an underlying rule or abstract relationship, even across modalities, but not when an overlap is purely perceptual (Garner et al., 2016)."

"At the same time, reuse of shared abstract representations introduces a computational trade-off, such that overlapping codes that facilitate transfer can also induce interference, as observed in both humans and artificial neural networks (Holton et al., 2025). However, while there is now much work elucidating how learned representations are generalized, it remains unclear whether and how abstracted structural representations are expressed neurally during ongoing stimulus processing, and how they are shaped by prior structural experience. Thus there is a need to characterize not only where abstract representations are located in the brain but also how they dynamically unfold and are reused across contexts in support of compositional generalization."

> *(2) I found it interesting that the authors chose to perform PCA for dimensionality reduction prior to conducting RSA; however, I haven't seen such an approach in the literature before. It would be helpful to either cite prior studies that have employed a similar method or to include a comparison with more standard approaches, such as sensor-level RSA or sensor-searchlight analysis.*
>
> *(3) Connected to the previous point, the choice to use absolute distance as a dissimilarity measure is not justified. How does it compare to standard metrics such as correlation distance or Mahalanobis distance? The same applies to the use of Kendall's tau.*

We thank the reviewer for raising these methodological concerns and agree methodological choices should be explicitly motivated and justified. Our primary motivation for using PCA, and retaining components that jointly explain 80% of the variance, was denoising. Sensor-level MEG data are high-dimensional and highly correlated, and representational analyses that operate directly on the full sensor space may be susceptible to noise amplification and overfitting, particularly when estimating distance measures across time. Retaining components up to a fixed variance threshold is a widely used approach, both in MEG decoding and RSA, to preserve the dominant signal structure while attenuating sensor-specific noise. Importantly, this procedure discards only the lowest-variance 20% of the signal and therefore does not impose strong constraints on the representational geometry.

To address the reviewer's concern regarding methodological choices and what this means for the conclusions drawn more directly, and to ensure our findings are not driven by specific analytic choices, we conducted additional robustness checks, availing of more "standard" approaches. Specifically, we performed the same RSA on localizer tasks but using sensor-level data without PCA, employing multiple distance metrics (including absolute distance/L1 norm and Euclidean distance/L2 norm). These similarity metrics are commonplace in RSA, providing linear, per-dimension differences between neural patterns (L1) or robust small outlier differences along individual dimensions compared to squared differences (Euclidean/L2), providing simple and interpretable similarity metrics.

These sensor-level analyses yielded highly similar temporal profiles and statistical effects to the PCA-based results, including the key contrasts reported in the main text. We now include these results in a new Supplementary Figure and reference them explicitly in the Results section.

In the revised manuscript, we clarify that PCA was used as a principled denoising step, not as a means of imposing structure on the data, and explicitly show that our conclusions are robust across analytic choices.

"Importantly, to verify that the chosen PCA approach did not impose artificial structure within the data, or introduce other potential confounds, we also conducted the above analyses on sensor level MEG data. Highly similar results were obtained, with consistent

results for different similarity metrics (absolute distance (L1 norm) and Euclidean distance (L2 norm), Supplementary Figure 3)), demonstrating that our results and conclusions are robust across analytic choices."

Methods

"In addition to the PCA-based feature analysis, we performed a sensor-level RSA. At each time point, we estimated condition-specific activation patterns using a GLM applied separately to each sensor (dummy-coded stimulus identity). RDMs were then computed across stimuli using mean absolute difference (L1 norm) as the primary distance metric. For completeness, we additionally evaluated Euclidean distances (L2 norm), which yielded qualitatively similar results (Supplementary Figure 3)."

Our choice of Kendall's tau rank correlation was motivated by its common use for linking neural and theoretical RDMs (e.g., Luyckx et al., 2019 eLife; Nili et al., 2014 PLoS CB). We have now added these citations to the manuscript.

> *(5) The analysis described in the "Abstract representation of dynamical roles in subprocesses" does not appear to convincingly test the stated prediction of a structural scaffolding account. The authors hypothesize that if structure and dynamics from prior experiences are repurposed, then stimuli occupying the same "dynamical roles" across different sequences should exhibit enhanced neural similarity. However, the analysis seems to focus on decoding transitions rather than directly assessing representational similarity. Rather, this approach may reflect shared temporal representation in the sequences without necessarily indicating that the neural system generalizes the abstract function or position of a stimulus within the graph. To truly demonstrate that the brain captures the dynamical role across different stimuli, it would be more appropriate to directly assess whether neural patterns evoked by stimuli, in the same temporal part of the sequence, with shared roles (but different visual identities) are more similar to each other than to those from different roles.*

We thank the reviewer for this careful and constructive critique and analytical suggestion. The reviewer is correct that in the bridge graph traversal, structural roles map onto fixed serial positions within a sequence, meaning that same-role stimulus pairs also tend to share the same temporal position in the sequence. This is a potential confound: neural similarity between same-role stimuli could in principle reflect similarity related to time-in-sequence rather than an abstract structural role per se.

The original role-based RSA does not fully resolve this concern, and we have now conducted, based on the reviewer's suggestion, an additional analysis that directly addresses it. Rather than comparing stimuli that share a structural role, we examine whether the neural representation of a single central node stimulus is similar across its two distinct serial positions within the prior and transfer learning sequences. Specifically, in the bridge graph traversal, each central node appears twice per sequence: the left central node occurs at steps 1 and 4, the right central node occurs at steps 5 and 8. We estimate neural patterns for each central node separately at its first and second occurrence, using occurrence-specific GLMs, and then compute cross-phase similarity between prior and transfer learning: the neural pattern evoked by a central node at its first occurrence in prior learning is compared to its pattern at the second occurrence in transfer learning, and vice versa.

For the 6-bridge prior group, whose prior learning also involved a bridge graph, central nodes were identified using the same bridge-template positions (steps 1/4 and 5/8) in both prior and transfer learning phases. For the 4-cycle prior group, whose prior learning involved a line graph rather than a bridge graph, central nodes in the prior phase were identified using degree-based role assignment (the two nodes with the highest degree that are not boundary or intermediate nodes), ordered by first appearance in the prior trial sequence.

Transfer learning central nodes were identified from the bridge graph template in both groups, since transfer learning always involved a bridge graph. The group comparison thus tests whether cross-phase similarity between central node representations is specific to participants whose prior experience included a bridge graph relative to participants whose prior experience involved a structurally different graph, for whom "central" nodes in prior and transfer are central in qualitatively different senses. This approach directly controls for the time-in-sequence confound, since the same node is compared across different serial positions rather than comparing different stimuli that share a role and therefore also share a sequence position.

We compared this cross-phase similarity between the 6-bridge prior and 4-cycle prior groups. We observe a significant group difference at a late stage of the trial (approximately 1.125–1.25 sec post-stimulus onset, $P^{FEW} = 0.043$, two-sample non-parametric permutation test, one-tailed), with additional differences at earlier time points that did not survive correction for multiple comparisons (0.125 sec: t = 2.18, p = .034; 0.79 sec: t = 2.84, p = .007). This indicates increased neural similarity for central node representations across phases in participants who experienced these roles consistently during prior learning.

We wish to be entirely transparent, however, about two limitations of this additional analysis with respect to the original dynamical role decoding results reported in Figure 4. First, the temporal profile of the cross-phase similarity effect does not perfectly overlap with the time windows identified in the original decoding analysis, which limits the strength of the convergence argument. We therefore cannot fully rule out a possibility that the earlier decoding effects partly reflect temporal position within the sequence rather than abstract structural role coding per se. Second, this analysis is necessarily based on a smaller and less variable set of comparisons than either the original decoding analysis or the role-based RSA, since the bridge graph contains only two central nodes whose repeated occurrences can be directly compared in this way. This limits the statistical power of the analysis and our ability to draw strong conclusions from it in isolation. We acknowledge both limitations explicitly in the revised manuscript. Nonetheless, the observation of a group-specific effect in the cross-phase analysis, which by design controls for serial position, provides some support for the structural scaffolding account, and the late corrected cluster suggests that role-consistent representations do emerge across learning phases.

In addition to this online analysis, we conducted a parallel role-based RSA on data from the localizer task. This localizer-based analysis examines representational changes from PRE to POST learning for stimuli that share the same structural role within the relevant sub-process graph structures. We find a cluster-based permutation-corrected interaction effect between prior learning group and graph factor, indicating that representational similarity increases selectively for stimuli sharing the same structural role, rather than uniformly for all stimuli associated with a given graph. Together with the localizer-based RSA results, we believe these converging analyses strengthen the overall interpretation, while we acknowledge the temporal confound as a remaining limitation that future work with designs explicitly counterbalancing serial position across roles could address more definitively. We reproduce below the main text results, supplementary methods and Supplementary Figure (Supplementary Figure 4) detailing this complimentary analysis.

Main manuscript

“To further test the robustness of the above effect, and to assess whether abstract dynamical roles are reflected in neural representational geometry during task performance, we conducted a complementary cross-phase representational similarity analysis. This analysis addresses directly whether neural representations of central nodes are stable across their distinct serial positions within a sequence, an expected property if representations reflect abstract structural identity rather than time-in-sequence. Specifically, for each participant we estimated neural patterns for each central node separately at its first and second occurrence

(phase) within the sequence traversal, using occurrence-specific GLMs, and computed cross-phase similarity between prior and transfer learning phases. This approach controls for a potential time-in-sequence confound inherent in role-based comparisons, since now, the same functional node is compared across different serial positions (rather than comparing different stimuli with a common sequence position, by virtue of sharing role). Consistent with the decoding results, we observed increased cross-phase similarity in participants with prior experience of Component 6 relative to those with prior experience of Component 4 (Supplementary Figure 4). Specifically, a significant group difference was observed at a late stage of the trial (1.125–1.25 sec post-stimulus onset, $P_{FWE} = 0.043$, two-sample non-parametric permutation test, one-tailed), with additional differences at earlier timepoints that did not survive correction for multiple comparisons (0.125 sec: t = 2.18, p = .034; 0.79 sec: t = 2.84, p = .007). These effects indicate increased neural similarity for central node representations across phases in participants who experienced these roles consistently during prior learning. We consider these results complementary to the main analyses and interpret them with caution given that the group difference was only corrected for multiple comparisons at the late cluster. Importantly, the temporal profile of the cross-phase similarity effect does not perfectly overlap with the time windows identified in the original dynamical role decoding analysis, which limits the strength of the convergence argument. We therefore cannot fully rule out the possibility that the earlier decoding effects reported in Figure 4 partly reflect temporal position within the sequence rather than abstract structural role coding per se. We additionally note that this analysis is based on a smaller and less variable set of comparisons than either the original decoding analysis or the role-based RSA, since the bridge graph contains only two central nodes whose repeated occurrences can be directly compared in this way, limiting its statistical power in isolation. Nonetheless, the observation of a group-specific effect in the cross-phase analysis which directly controls for serial position provides support for the structural scaffolding account, and the late corrected cluster suggests that role-consistent representations do emerge across learning phases, albeit potentially at a later stage of processing than the decoding results alone would suggest."

Supplementary Information

"Cross-phase representational similarity analysis

Complementary to the decoding analysis in Figure 4, we conducted a cross-phase representational similarity analysis (RSA) to test whether neural representations of central nodes were stable across their distinct serial positions within a sequence, a property expected under abstract structural role coding but not under serial position coding. We focused on central nodes of the 6-state graph, as these nodes appear twice within each sequence traversal of the bridge graph: the left central node occurs at steps 1 and 4, the right central node occurs at steps 5 and 8, allowing first and second presentations to be dissociated.

Structural roles were assigned separately for prior and transfer learning phases. For line graphs (experienced by the Condition 4 during prior learning), central nodes were identified from node degree in the adjacency matrix (boundary: degree 1; intermediate: neighbors of boundary; central: remaining nodes), and the two central nodes were ordered by first appearance in the trial list. For bridge graphs, central nodes were identified from the predefined 8-step traversal template.

At each time point and sensor, condition-specific activations were estimated using occurrence-specific GLMs: one GLM fitted to trials corresponding to the first occurrence of each number in the trial list, and one fitted to trials corresponding to the second occurrence. Component 6-specific beta estimates were extracted from each GLM separately, yielding four sets of neural patterns per subject: prior-phase first occurrence, prior-phase second occurrence, transfer-phase first occurrence, and transfer-phase second occurrence.

Cross-phase representational dissimilarity was then computed using an L1 distance metric (mean absolute difference across sensors) across four cross-presentation pairs: Left central (prior first, transfer second), left central (prior second, transfer first), right central (prior first, transfer second), and right central (prior second, transfer first). These pairings compare the same node across different serial positions and different learning phases, directly controlling for time-in-sequence as a potential confound. The mean L1 distance across the four pairs constituted the cross-presentation dissimilarity index at each time point, where lower values indicate greater neural similarity. Timecourses were baseline-corrected by subtracting the pre-stimulus mean (timepoints 1–25, corresponding to −250 to 0 ms) per participant, and analyses were restricted to post-stimulus timepoints. Group differences between Condition 6 and Condition 4 participants were assessed using a two-sample non-parametric cluster permutation test (cluster-forming threshold p = 0.10, cluster-level alpha = 0.05, 1,000 permutations, one-tailed, consistent with the directional prediction that Condition 6 participants would show greater cross-phase neural similarity for central nodes). The results are presented in Supplementary Figure 4."

> *(4) In the following section, the authors correlate decoding accuracy with participants' behavioral performance across different conditions. However, out of the four reported correlations and the additional comparison of differences between conditions, only one correlation and one correlation difference reach significance, and only marginally so. The interpretation of this finding should therefore be more cautious, especially if it is used to support a link between neural representations and behavior. Additionally, it is possible that correlation with a more clearly defined or targeted neural signature, more directly tied to the hypothesized representational content, could yield stronger or more interpretable correlations.*

We agree with the reviewer that the reported brain-behavior correlations should be interpreted cautiously. Only one of the four correlations reached significance, with effect sizes of medium magnitude. However, we emphasize this correlation is significant when corrected for multiple comparisons for two tests, experience and inference probes (α = 0.025). Additionally, it is important to note, we did not expect all four correlations to be significant. The conditions differ in the extent to which behavior is predicted to depend on the specific neural representation indexed by the decoding analysis, and only a subset of conditions are theoretically expected to show a brain–behavior relationship. The absence of significant correlations in the remaining conditions is therefore not inconsistent with our account. For example, we did not expect the two 4-cycle correlations (depicted in green, Fig. 4C) to be significant.

We agree with the reviewer that the reported brain-behavior correlations should be interpreted cautiously, and have revised the manuscript to reflect a more measured interpretation. We note that the correlation value reported in the revised manuscript (ρ = .48, p = .016) differs slightly from the originally submitted version (ρ = .41, p = .041), due to a correction in the timepoint-averaging window used to derive the decoding strength measure. We consider the revised value the authoritative one and have updated all relevant text accordingly.

Only one of the four correlations reached significance, with an effect size of medium magnitude. We would not characterise this correlation as marginal: it survives correction for multiple comparisons across the two theoretically motivated tests (experience and inference probes; corrected α = 0.025), and a medium-magnitude Spearman correlation (ρ = .48) in an individual-differences analysis with n = 25 is consistent with a meaningful brain-behavior relationship, even if the sample size limits the precision of the estimate. We agree that independent replication is needed to establish robustness.

Additionally, it is important to note that we did not expect all four correlations to be significant. The conditions differ in the extent to which behavior is predicted to depend on the specific neural representation indexed by the decoding analysis, and only a subset of conditions are theoretically expected to show a brain–behavior relationship. The absence of significant correlations in the remaining conditions is therefore not inconsistent with our account. For example, we did not expect the two 4-cycle correlations (depicted in green, Fig. 4C) to be significant, since the 4-cycle-prior group did not experience the bridge graph structure during prior learning and therefore would not be expected to show the same alignment between dynamical role representations and transfer learning performance.

In any event, based on the reviewer's (and comments from other reviewers about these correlations), we have revised the manuscript to tone down the interpretation of these results (reproduced below):

"These findings are consistent with, though not conclusive evidence for, dynamical role representations conveying behavioral benefits. Speculatively, such knowledge representation might enable individuals to constrain their predictions of 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". While we acknowledge that the correlation results appear inconsistent across experience and inference probes, it is important to consider that our MEG classifiers were trained and tested on experienced transitions alone, such that enhanced classification accuracy is more reflective of a neural representation of experienced transitions. Under this assumption, the MEG classifier would more closely align with behavioral accuracy for the experience probes (testing actually experienced transitions) than for the inference probes (testing never experienced but only implied transitions on the graph). See Discussion for further elaboration of this point, including why this asymmetry is expected under a model of gradual, imperfect structural learning.

The observed correlation is also consistent with a general learning-ability account, whereby higher-ability participants both better discover the Component 6 structure (that happens to be shared) across sessions and show higher behavioral accuracy overall. For example, higher-ability participants may rediscover this structure during transfer, without necessarily benefitting from having been exposed to it during primary learning. This ability-based account predicts that decoding strength should also relate to Component 4 accuracy, and that Component 4 and Component 6 accuracy should correlate within each prior learning group. Neither prediction was consistently supported (Supplementary Figure 5): The Component 4 decoding strength and Component 4 correlation was not significant ($\rho = .38$, $p = .071$), and did not differ significantly from the Component 4 decoding strength and Component 6 correlation (Steiger's test, $p = .656$); accuracy correlations between Component 4 and Component 6 were significant in only one prior learning group for each probe type, with no significant between-group differences (all $p > .29$).

Given this inconsistent pattern, we consider the general-ability account unsupported, though our low per-condition sample size limits power to rule it out. We therefore treat the brain-behavior correlation as selective to Component 6 experience probes in this sample, and treat it as supportive rather than primary evidence for a link between role representations and transfer, pending independent replication."

We agree that a more targeted neural signature could in principle yield a more interpretable correlation with behavior. Our interpretation of the reviewer's comment is that this implicitly requires a measure that captures representational content at the level of individual trials or items, for example, a trial-by-trial estimate of how strongly a given stimulus is represented in its abstract structural role, which could then be correlated with individual participants' behavioral performance. Our current decoding measure does not provide this: it indexes classification accuracy averaged across all trials of a given transition type, yielding a single

scalar value per participant per timepoint rather than a trial- or item-level representational estimate.

Computing such a trial-level measure from our data would require a fundamentally different analysis approach, for example, single-trial decoding confidence scores or RSA-based measures of individual stimulus representational fidelity, which goes beyond what our current classifier-based pipeline and data structure can deliver. We have instead conducted additional analyses that more directly index neural representational content (role-based representational similarity analyses, reported in the revised manuscript), which provide converging support for the proposed computational account.

> *Minor Comments:*
>
> *"During preprocessing, sensors were excluded based on an identified noise level. However, the authors do not specify the threshold used to define this noise level, nor do they report how many sensors were excluded per participant. It would be helpful to have these details. Additionally, it is unclear why the authors opted to exclude sensors rather than removing noise with MaxFiltering or interpolating bad sensors. Finally, the authors should report how many trials were discarded on average (and standard deviation) per participant."*

We thank the reviewer for these comments. We have updated the methods section to address the points raised.

Regarding the noise threshold and number of excluded sensors/trials: sensors were excluded using the default osl_detect_artefacts settings, which apply a Generalised Extreme Studentised Deviate (GESD) test at a significance level of 0.05. On average across all MEG runs and participants, 8.67 sensors (SD = 7.36) and 9.13 trials (SD = 11.46) were excluded. These details have been added to the methods section.

Regarding the choice to exclude sensors rather than applying MaxFilter or interpolation: MaxFilter was not applicable here as our data were acquired on a CTF system, for which MaxFiltering is not designed. Sensor interpolation was not performed as bad channels are excluded prior to ICA decomposition following standard OSL pipeline practice. Given that excluded sensors represented on average ~3% of the total 273 sensors, we do not expect this to have meaningfully affected our results.

The updated methods section now reads as follows:

"On average across all MEG runs and participants, 8.67 sensors (SD = 7.36) and 9.13 trials (SD = 11.46) were excluded (using default osl_detect_artefacts settings; channel significance level for Generalised Extreme Studentised Deviate = 0.05). We did not apply MaxFilter as our data were acquired on a CTF system, for which MaxFiltering is not designed. Sensor interpolation was not performed as bad channels are excluded prior to ICA decomposition following standard OSL pipeline practice; given that excluded sensors represented on average ~3% of the total 273 sensors, we do not expect this to have meaningfully affected the results."

> ***Reviewer #3 (Recommendations for the authors):***
>
> *(1) The manuscript seems to use the terms generalization, transfer, compositionality, and compositional generalization somewhat interchangeably. While I appreciate that both compositionality and transfer learning are central to the task design, I believe it's important to distinguish these concepts more clearly. Specifically, generalization typically refers to the ability to apply learned structure within a task to novel but similar instances, whereas transfer involves applying learned components or representations across different tasks or contexts. Clarifying these distinctions - ideally with explicit definitions*

> *early in the introduction - would help avoid confusion and better situate the contribution of this work within the broader literature.*
>
> *(2) The manuscript frequently refers to abstract knowledge or abstract representations. However, I would argue that abstract knowledge is typically characterized by its ability to generalize or transfer across different tasks or domains. In the current paradigm, while the structural relationships between stimuli change, the task demands and sensory modalities remain constant. It would strengthen the conceptual clarity of the paper if the authors could more carefully define what they mean by "abstract".*

We thank the reviewer for these suggestions. We agree that “abstract” can carry different meanings across the literature. We have now added an explicit definition to the introduction clarifying our usage. We use 'abstract' to refer to representations that are invariant to surface sensory features while capturing relational structure, consistent with findings in a representation learning literature (e.g., Bernardi et al., 2020; Behrens et al., 2018). While the reviewer raises the valid point that task demands and sensory modality remain constant across learning phases, we highlight that the stimuli themselves are entirely novel in the transfer phase, such that any representational similarity must reflect relational rather than perceptual overlap.

We have added explicit definitions of abstract, generalization, transfer, and compositional generalization to the introduction, clarifying how each term applies to our task design and how they relate to one another. We have also aimed for more consistent usage throughout the manuscript.

This added paragraph in the Introduction reads as follows:

“Throughout this work, we use "generalization" to refer to the ability to apply learned relational structure to novel stimulus compositions within the same task context, and "transfer" to refer to the reuse of (abstracted) representations across distinct task contexts or learning phases. Note we use “abstract” to refer to representations that are invariant to the sensory specifics of individual stimuli, capturing relational structure (e.g., the role a stimulus occupies within a graph) rather than surface features. Our task is structured compositionally in that every experienced compound stimulus is the joint product of two simultaneously evolving subprocesses, while the specific compound encountered during transfer learning is never observed during prior learning. We use 'decomposition' and 'transfer' to refer to the abstraction and reuse of a single such subprocess across learning phases, and reserve “compositional generalization” for the broader phenomenon whereby task performance depends on appropriately decomposing and recombining structural knowledge, a capacity our design requires of participants. Here we acknowledge that our study speaks most directly to decomposition and transfer of individual structural components rather than to the act of recombination itself.”

> *(3) The supplementary materials include a section on replay analyses, but it is unclear how these analyses relate to the main text. I could not find a corresponding discussion or reference to replay in the results or discussion sections.*

We thank the reviewer for flagging this. The replay analysis was referenced in the Discussion of the previous version of the manuscript. To make this analysis more visible, we have also added a brief summary of the null replay result to the Results section, with a pointer to the Supplementary Results where the full analysis and discussion of limiting factors are reported.

https://doi.org/10.7554/eLife.107162.2.sa0