communications psychology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s44271-025-00259-w

Feature identification learning both shapes and is shaped by spatial object-similarity representations

Jonathan K. Doyon [1,2,3] ✉, Sarah Shomstein [1] & Gabriela Rosenblau [1,2] ✉

Object knowledge is bound together in semantic networks that can be spatially represented. How these knowledge representations shape and are in turn shaped by learning remains unclear. Here, we directly examined how object similarity representations impact implicit learning of feature dimensions and how learning, in turn, influences these representations. In a pre-experiment, 237 adult participants arranged object-pictures in a spatial arena, revealing semantic relatedness of everyday objects across categories: activity, fashion, and foods. The subsequent experiment assessed whether these semantic relationships played a role in implicitly learning specific object features in a separate adult participant group (N = 82). Participants inferred the meanings of two pseudo-words through feedback. Using computational modeling, we tested various learning strategies and established that learning was guided by semantic relationships quantified in the pre-experiment. Post-learning arrangements reflected object similarity representations as well as the learned feature. We directly show that similarity representations guide implicit learning and that learning in turn reshapes existing knowledge representations.

Humans are shaped by experience, yet how these experiences dynamically aggregate into knowledge structures to guide future experiences is not well understood. How do we acquire knowledge, and how does this knowledge, in turn, shape future learning? One intriguing notion is that the brain's primary function is to extract statistical structure from discrete experiences with the extracted structure forming the basis for learning[1–5]. This process, coined implicit learning, has been investigated across domains ranging from perceptual decision making[6,7] to social learning[8]. While studies have shown the aggregation of rules through repeated interactions with the environment, it remains unclear how new rules are integrated into pre-existing knowledge representations, such as knowledge of how objects relate to one another or category belonging.

Preexisting knowledge structures often coined cognitive schemata[9–12] have been mostly investigated in category or prototype learning. Typically, studies investigated a small and discrete set of semantic categories and showed that these semantic structures group items and facilitate the deployment of attention and learning across perceptual decision-making tasks[13,14]. Conversely, having to group items that are semantically unrelated or incongruent is more effortful and error prone[15]. While this literature establishes the importance of semantic categories for human perception and cognition, newer studies show that semantic knowledge structures may be finer-grained and more flexible.

Fine-grained semantic knowledge in the form of similarities or transition probabilities between events has been shown to play an important role in visual perception[3,16], learning abstract information[17,18], and language processing[19,20] as well as in social inferences[21–23]. Spatial representations of semantic relationships, captured though clever semantic relatedness rating tasks, have been shown to correspond to patterns of brain activity[24–28]. These findings corroborate that the brain encodes the multivariate statistical structure of object configurations, from fine-grained similarities between objects to coarser feature representations.

While a substantial body of work has focused on the structure of conceptual knowledge, such as semantic relatedness across various cognitive domains, less is known about how pre-existing conceptual knowledge shapes and is in turn shaped by active learning. In the perceptual attention and priming literature, studies have shown that the similarity of objects influences categorization performance and that vice versa, perceived object categories can change with selective attention to a feature in question[29–31]. The current study examined whether object similarity guided learning in the reinforcement learning (RL) framework, specifically, and whether, in turn,

[1]Department of Psychological and Brain Sciences, The George Washington University, Washington, D.C., USA. [2]Autism and Neurodevelopmental Disorders Institute, The George Washington University, Washington, D.C., USA. [3]Schepens Eye Research Institute of Massachusetts Eye and Ear, Department of Ophthalmology, Harvard Medical School, Boston, MA, USA. ✉e-mail: jdoyon1@meei.harvard.edu; grosenblau@gwu.edu

learning to correctly identify a feature based on task feedback influenced the representation of object similarity.

In the reinforcement learning (RL) framework, an agent learns through environmental feedback by updating prediction errors (PEs)—the difference between expected and received feedback. RL models constitute robust algorithms that characterize learning processes across a wide range of tasks on the behavioral and neural levels[32–35]. The simplest RL rule, Rescorla–Wagner learning, posits that agents acquire information through trial-and-error. This simple, model-free rule requires fewer resources but can be quite inefficient because the agent has to experience all possible states to learn from them[36].

A more efficient but computationally taxing form of RL, model-based learning, assumes that a learner abstracts from the learned material to the underlying task structure. This allows the learner to flexibly generalize knowledge to unseen items or situations[32,36]. For example, in a study by Kahnt and colleagues, people generalized task feedback across similar features[18]. A simpler formalization of task structure is scaling model free PEs based on the similarity of stimuli, for instance-, similarities between faces[8] or between mental states (e.g., inferred from traits and preferences for items). In previous studies, we have demonstrated that adults and adolescents apply similarity-based PE updating to learn about other people's traits and preferences[21,22,37]. It is an open question whether this type of similarity learning generalizes across cognitive domains.

Recent advances have allowed studies investigating semantic similarity to rely on multidimensional scaling tasks, such as the multi-arrangement task (MAT), to establish fine-grained representational dissimilarity matrices (RDMs, e.g.,[38]. In the MAT, item (dis)similarities are captured spatially, by asking participants to place similar items close to one another, and dissimilar items proportionately further apart, thereby using spatial distance as a measure of semantic relatedness. The resulting RDMs convey the semantic relationships between all item-pairs in a task set and can be reduced to the core underlying dimension(s).

The current study asks an important question: to what degree object semantic similarity (quantified by spatial maps derived from the MAT) guides implicit feature identification learning (i.e., mapping nonwords to specific predefined item features). By integrating computational modeling of learning task behavior and comparing direct assessments of semantic relatedness in the absence of learning and after learning, this study directly examines the influence of semantic relatedness on learning and the dynamic shifts in semantic relatedness induced by the learning process.

In the pre-experiment, we investigated the object semantic similarity structure with the MAT task. In the main experiment, we tested whether and how a separate group of participants used object-to-object semantic similarity structure, established in the pre-experiment, to implicitly learn about word meanings. Specifically, participants were asked to map a non-word to a specific feature (i.e., how colorful or large objects are). They could learn about the meaning of the nonword (i.e., the feature in question) through trial-by-trial feedback. The same objects were used in both experiments; therefore, we could directly assess the degree to which the semantic relationships among real-world objects assessed in the pre-experiment shifted as a function of learning in the main experiment. To directly test whether participants relied on object semantic similarity during learning, we used a computational modeling framework. Based on the previous literature, we predicted that learning activates semantic similarity maps as they represent important conceptual knowledge. Lastly, we predicted that learning, in turn, changes these preexisting knowledge representations. To this end, we compared the object semantic similarity maps after learning in the main experiment to those that were established in the absence of learning in the pre-experiment.

Methods

Participants

Four hundred young adults (age range: 18–25 years, 300 for the pre-experiment and 100 for the main experiment) were recruited from the online Prolific research participation platform (https://www.prolific.co/). Sixty-three participants began but did not complete the experimental tasks and were excluded (55 in the pre-experiment and 8 in the main experiment). Further exclusions were made due to excessive missing data (more than two standard deviations above the average number of missing trials, 8 in the pre-experiment and 10 in the main experiment).

The final sample size for the pre-experiment was 237 adults (126 identified as female participants, 109 identified as male participants, and 2 preferred no response; mean age = 22.26 years, SD = 2.24).

The final sample for the main experiment comprised 82 adults (61 identified as female participants and 21 identified as male participants; mean age = 21.61 years, SD = 2.02). The sample size rationale, exclusion criteria, and missing responses criteria can be found in the supplement. Participants provided informed consent before participating. The study was approved by the Institutional Review Board (IRB#: NCR191133) and was conducted in accordance with the Declaration of Helsinki. Participants were compensated $10 USD per hour of participation (pre-experiment mean duration = 59.64 min, SD = 22.42; main experiment mean duration = 67.10 min, SD = 21.01).

Pre-experiment—inferring spatial similarity representations via multidimensional scaling

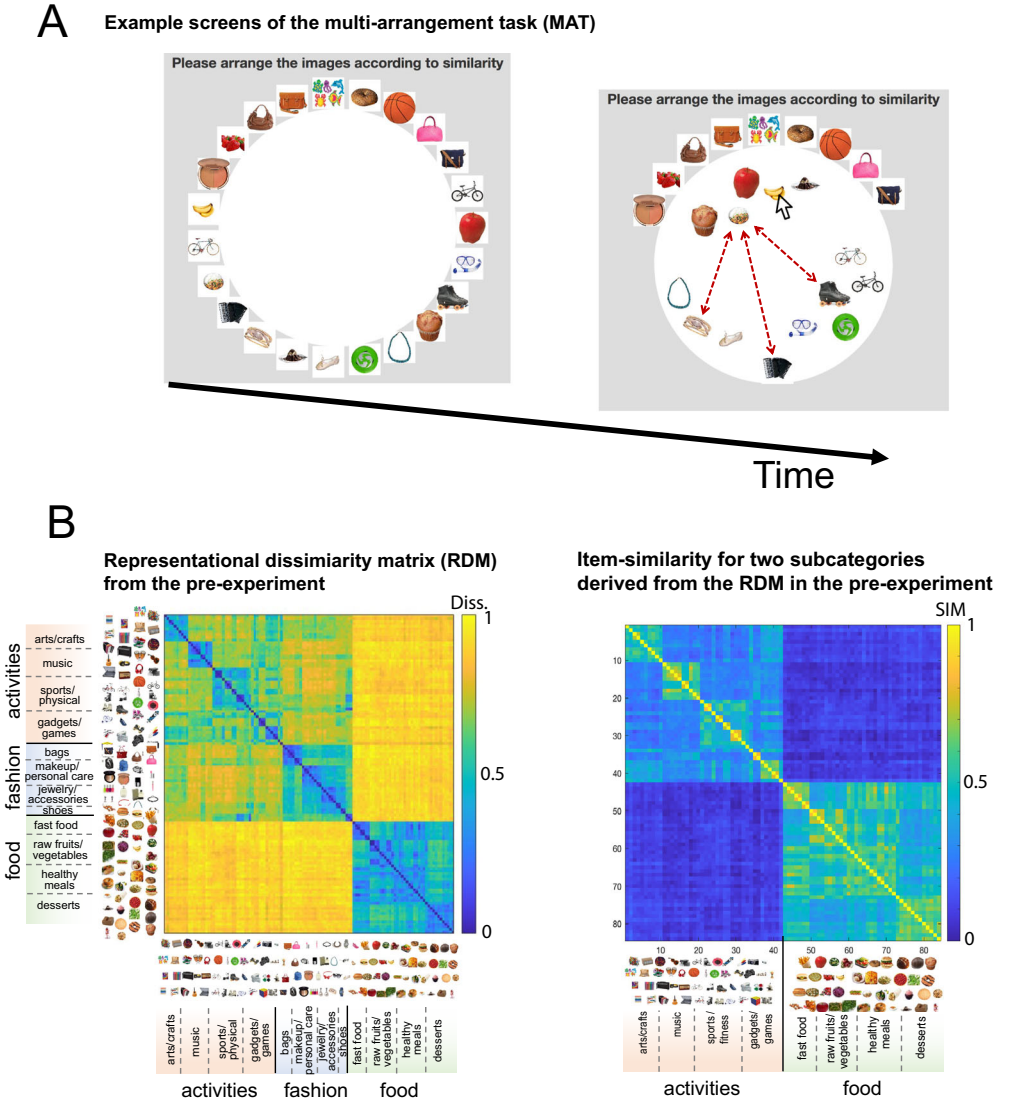
The pre-experiment used a set of 120 images of objects belonging to one of three categories, each of which contained four subcategories: 49 activity items (arts and crafts; music; sports; toys, gadgets, and games), 29 fashion items (accessories; bags; cosmetics; shoes), and 42 food items (fast food; healthy savory food; raw fruits and vegetables; sweets; see supplement for stimuli details). To capture and quantify semantic relationships among objects (e.g., an apple is more associated with a cake than a purse), we employed multidimensional scaling. In an online study, participants performed the MAT (see supplement for a detailed task description and experimental design). In brief, objects were positioned around a circular arena (Fig. 1A) and participants were asked to place each object into the arena according to how similar each pictured object was to one another (i.e., similar objects should be placed near one another, and the position of any given object represents its similarity/dissimilarity to every other object present in the arena). Some study materials are publicly available (https://meadows-research.com/documentation/researcher/tasks/multiple-arrangement). After the task, they rated how colorful, expensive, and large the items were (see supplement for details on the item rating task). No aspects of the study were preregistered.

Statistical analysis

To investigate how participants represented item-level semantic relationships, we first used a validated multidimensional scaling approach[38,39] yielding multidimensional representations of object dissimilarities in the form of representational dissimilarity matrices (RDMs; Fig. 1B, left, see supplement for a description of the item-level dissimilarities). The RDM entailed the average dissimilarity of each object to all others in the set ($n = 120$ averaged dissimilarities) for each participant ($N = 237$ participants or observations). In order to test whether object arrangements entailed the *a priori* category structure, we fit two linear models to predict the z-scored averaged dissimilarities (DIS) with category and subcategory information. Here, and in subsequent models, statistical assumptions were checked prior to each analysis. We used generalized least-squares estimators to yield unbiased estimates where data were non-normal and weighted versions when variances were not constant.

$$\mathrm{DIS} \sim \mathrm{category} + \varepsilon \quad (1)$$

$$\mathrm{DIS} \sim \mathrm{subcategory} + \varepsilon \quad (2)$$



We performed a PCA analysis to validate the a priori defined coarse semantic structure of the item set (i.e., category and subcategory assignment of items). To this end, we tested whether the a priori defined object categories and subcategories could be recovered from the PCA components, and how many principal components were needed for this recovery. To do this, we fit a binomial generalized linear model[40] and predicted item-pair category and subcategory membership using the data reconstructed with the first principal component (k_1), then the first two components ($k_1 + k_2$), and so on ($k_1 + k_2 \ldots + k_k$). We selected the number of components k that significantly improved the model by comparing each simpler model with k_n components to the one with k_{n+1} components with the ANOVA function in R. This function tests whether the more complex model is significantly

Fig. 1 | Overview of the pre-experiment. A Sample multi-arrangement task (MAT) trial. Objects ($n = 120$) begin at the periphery of the arena (left). Participants ($N = 237$) then spatially arrange objects by using a mouse (right). Red arrows (not depicted during the experiment) indicate dissimilarities measured by the Euclidean distances between objects' positions. **B** Left panel: Average representational dissimilarity matrix (RDM) obtained from the pre-experiment. Objects in their *a priori* category assignment are depicted along the axes. Each cell corresponds to the dissimilarity (Diss.) between the intersecting objects. The diagonal identity line represents zero dissimilarity. Right panel: Object similarity matrix for activity and food items only. This was obtained by first computing item similarity from the dissimilarity values. Second, values were rescaled into the 0 (most dissimilar) to 1 (most similar) range. This similarity matrix was used in the computational models in the main experiment. **C** Results of the principal components analysis (PCA). RDM reconstructed using the first principal component (left) and the first two components (right) recover the object categories. The first two components of the PCA recover category and subcategory information. The first component differentiates food items from non-food items, while the second component differentiates activity and fashion items.

better at capturing the data than the simpler model.

$$\text{category} \sim k_1 + k_2 \ldots + k_n + \varepsilon \quad (3)$$

$$\text{subcategory} \sim k_1 + k_2 \ldots + k_n + \varepsilon \quad (4)$$

Category and subcategory predictors were coded as binary variables with 1 corresponding to item-pairs belonging to same categories and 0 to different categories.

Another aim of the PCA was to identify the key dimensions underlying item placements in the MAT. We hypothesized that participants' similarity ratings would reflect the a priori defined semantic category structure, while finer-grained item relationships across categories would also account for significant variance in item placements. We hypothesized that successful learning not only recruits coarse category representations but rather item-to-item relationships.

Finally, our goal was to probe whether the two selected features for the learning task in the main experiment were among the most relevant dimensions for item arrangements. We wanted to ensure that the learning task dimensions were not among the main dimensions. This would make sure that participants would have to learn the task, shifting representations from semantic knowledge (e.g., category, preferences, etc.) to the feature in question based on trial-by-trial feedback.

Main experiment—modulation of spatial similarity representations via implicit concept learning

Having established and quantified the components of the semantic space for our real-world items, the main experiment tested our main hypothesis that knowledge structures, here spatial representations of object similarity, are flexibly deployed during feature identification learning and are, in turn, updated as a result of learning. We surmised that this process of deploying and updating knowledge constitutes a basic principle of how knowledge structures are formed and refined across cognitive contexts.

In an online study, we asked participants to first complete an implicit feature identification learning task (in short: feature identification task) involving a subset of the items used in the pre-experiment. In this task, a new set of participants was presented with a non-word, which corresponded to a specific object feature (e.g., how colorful objects are). Participants could learn about the respective word meaning and object feature through trial-by-trial feedback. Note that implicit learning in this context was defined as extracting regularities from situations without verbal explanations or rules[4,5]. Our task was inspired by implicit language acquisition tasks, in which learners engage with an unknown language without being provided with the grammar rules or asked to attend to these rules[41].

The feature identification task introduced in this study represents a generalization of the previously introduced social learning framework[21,22]. This task will be used in the context of a larger project on social and non-social learning, as detailed in our preregistration (https://osf.io/wvb8n). The task procedures and computational modeling approach used here, along with our hypotheses for the non-social learning task, were pre-registered before analyzing the data. The analysis plan and current sample were not preregistered.

After the feature identification task, participants completed the MAT, which was also used in the pre-experiment (experimental design procedures are described in the supplement). We tested participants' learning strategies via computational modeling and also how object similarity representations changed during learning. If representations of similarity were dynamically updated following learning, then representations generated post-learning should differ from those established outside of the learning context. To this end, the averaged RDM generated in the pre-experiment served as a baseline object similarity structure. This baseline RDM was then compared to the post-learning averaged RDM from the main experiment. We hypothesized that (1) participants learn about the feature in question, (2) item-level similarity representations guide learning during the task, and (3) the features that participants learned about are reflected in their post-learning item-level representations.

Feature identification task. Participants were introduced to a robot-language word (either "soaf" or "ation"), and asked to rate 42 objects per task run according to how "soaf" or "ation" they believed the depicted object to be (Fig. 2A). Unknown to the participant, the non-words either indicated how *colorful* or *large* the objects were. The association of non-words to the two features was counterbalanced across participants. One-half of the participants were presented with the word-feature combinations soaf-colorful and ation-large, while the other half of the participants were presented with ation-colorful and soaf-large. We chose the two features, colorful and large, that were not significantly related to the main semantic features of the object similarity space and also not significantly related to each other ($r = -0.138$, $p = 0.21$). This was done for two reasons: (1) we wanted to test whether participants represented semantic structure during learning even when the features in question were largely independent from the semantic relatedness structure and (2) in order to test whether learning induced a shift in the semantic similarity structure towards the learned feature, it was important for this feature to be independent, i.e., orthogonal to the main dimensions of the feature space.

Participants were instructed to use a Likert scale ranging from 1 "not at all" to 10 "very much" to indicate the extent to which the non-word (and assigned feature) applied to each object. After providing a rating, the participant received feedback about how much the non-word and assigned feature actually applied. The feedback that participants received was the averaged feature ratings from the pre-experiment (e.g., how colorful did participants in the pre-experiment rate this object on average). To add some variability in feedback given and ensure the use of a wider rating scale range, normally distributed random noise (one-half standard deviation) was added or subtracted from the average feedback ratings. The feedback values were then rounded to the nearest integer.

Participants completed two runs of the learning task. Participants either rated how colorful (e.g., soaf) or large (e.g., ation) the objects were. Run order (i.e., colorful or large first) and word-to-concept mapping (i.e., soaf meaning colorful or large) were counterbalanced across participants. No item was repeated in the task. In each run, participants completed 42 trials per run (84 total trials and items across the two runs). After completing each learning run, participants completed the MAT (see supplement for details). There were significantly more female than male participants in the final sample of the main experiment ($\chi^2(2) = 54.69$, $p < 0.001$). Concept learning and item-similarity arrangements did not differ significantly between male and female participants (learning task performance as explained by sex: $st.\beta = 0.02$, SE = 0.10, $p = 0.58$, 95% CI = [−0.13, 0.18]; sorting task performance as explained by sex: $st.\beta = -0.001$, SE = 0.0002, $p = 0.54$, 95% CI = [−0.007, 0.004]).

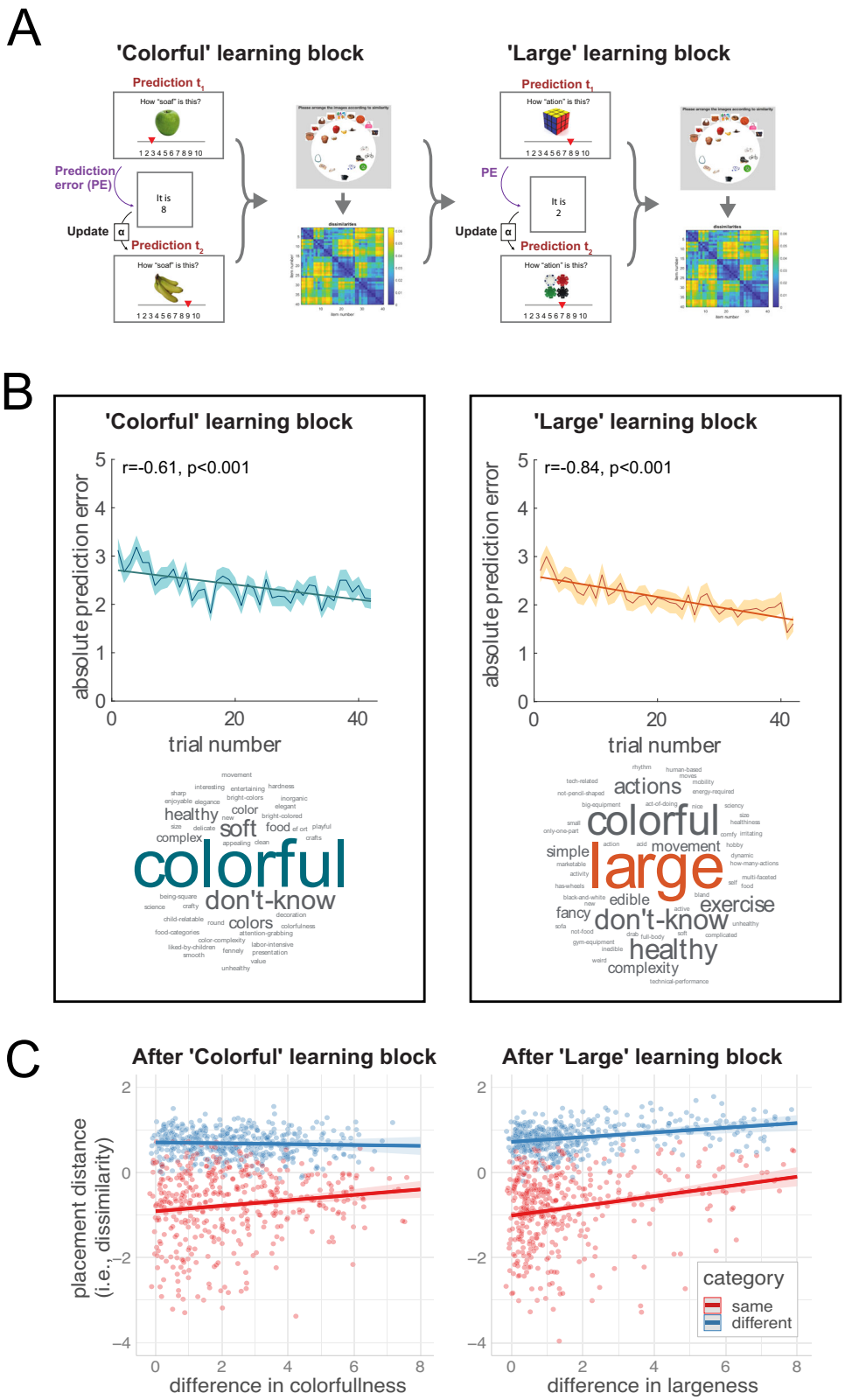


Fig. 2 | Overview of the main experiment. **A** Schematic trial sequence from the feature identification task. Participants ($N = 82$) rated the meanings of pseudo-words (e.g., "soaf") based on how much they apply to objects. The Rescorla–Wagner updating rule postulates that individuals update their estimates of the word in question based on the trial-by-trial feedback. This updating is leveraged by incorporating prediction errors (PEs), i.e., the difference between initial ratings and feedback, into subsequent ratings. The learning rate α is a free parameter, which captures the speed of learning (higher learning rates correspond to a faster integration of PEs). Following the learning task run, participants completed the multi-arrangement task (MAT). Participants completed two learning task runs, each followed by the MAT. **B** Upper panel: task-based PEs significantly decreased over trials in both task runs, evidenced by a negative correlation between PEs and trial numbers. Shaded regions indicate ±1 standard error. Lower panel: Participants' notions of the non-word meanings in response to open-ended questions about the non-words after each learning run. **C** Dissimilarities in colorfulness and largeness of object pairs ($n = 861$) predict object placement dissimilarities for objects of the same versus the different category. For instance, after the 'large' learning block, participants place objects of the same category (i.e., activities or foods) closer together if they are similarly large.

Statistical analysis

Feature identification learning. Participants were expected to learn about the feature in question. Here, learning is defined as a significant reduction in task-based prediction errors (PEs, i.e., the difference between participants' ratings and subsequent feedback, over the course of a task run. To this end, we computed bivariate correlations between PEs and the trial number for each of the learning runs.

The sample of the main experiment consisted of more female than male participants. There were no significant sex differences in PEs across conditions and no significant interactions between sex and feature condition (sex differences in PEs: $st.\beta = 0.02$, SE = 0.10, $p = 0.58$, 95% CI = [−0.13, 0.18]; sex by condition interaction: $st.\beta = -0.04$, SE = 0.15, $p = 0.63$, 95% CI = [−0.19, 0.12]). There were also no significant differences in learning (sex by trial interaction: $st.\beta = -0.01$, SE = 0.004, $p = 0.31$, 95% CI = [−0.04,

0.01]) or in post-task item pair arrangements (sex differences in post arrangements: $st.\beta = -0.002$, SE = 0.0001, $p = 0.41$, 95% CI = [−0.007, 0.003]). We therefore aggregated across male and female participants for the following analyses.

Learning-related changes to object similarity structures. Following each learning run, we assessed how participants perceived the similarity of the learning items. For this purpose, participants completed the MAT used in the pre-experiment with the same subset of items seen in the immediately preceding learning run. We then computed individual and average RDMs for the subset of items used in the two separate learning blocks. The procedure was identical to that reported in the pre-experiment (see Fig. 2A).

To assess how learning about the feature in question changed object similarity perception post-learning, we computed a feature dissimilarity regressor and assessed its effect on object placement. Feature dissimilarity (DIS) of two items *i* and *j* was defined based on the relationships between the feature ratings for these items (i.e. the correlation coefficient $\rho_{i,j}$ indicating how similar the objects *i* and *j* were with respect to how colorful and large they were rated by an independent sample in the pre-experiment).

$$\mathrm{DIS}_{i,j} = 1 - \rho_{i,j}$$

Assessing differences in object dissimilarity structures based on learning task features

To investigate whether the learning task influenced post-learning item-dissimilarity representations, two separate general linear models (GLMs) were constructed. Model 1 predicted item-pair dissimilarities post-learning based on item category, and additionally feature dissimilarity.

$$\mathrm{DIS} \sim \mathrm{category} \times \mathrm{feature} + \varepsilon \quad (5)$$

Model 2 predicted the item dissimilarity representations in the main experiment with the dissimilarity representations from the pre-experiment. To make the MAT dissimilarities comparable between the two experiments, we first created relative values by subtracting the average dissimilarity from item-level dissimilarities in each experiment. We included the additional predictors item category (same/different), item feature dissimilarity (i.e., the dissimilarity of colorfulness/largeness between item-pairs in respective colorful or large task blocks), and the interaction between these two predictors.

$$\mathrm{DIS}_{\mathrm{exp}\,2} \sim \mathrm{DIS}_{\mathrm{exp}\,1} \times \mathrm{category} \times \mathrm{feature} + \varepsilon \quad (6)$$

Computational modeling

To investigate how participants learned the meaning of the non-words, we employed the use of computational modeling. In recent work[21,22], we developed a computational modeling framework to describe how people learned about social information such as other people's preferences and character traits. These models use standard Rescorla-Wagner learning and or prior knowledge about the concept at hand. This framework can be applied to implicitly learning the non-word meaning by updating initial expectations about the feature in question through trial-by-trial feedback. Similar to social learning, we surmised that participants would learn the meaning of the non-word (i.e., learning about the feature in question) through trial-by-trial feedback scaled by item (dis)similarity.

Computational models. We introduced models of varying complexities, from simple regressions that capture a direct feature to non-word mapping, to more sophisticated hybrid Rescorla-Wagner learning models with additional knowledge about the feature in question. The computational models are described in detail in the supplement.

The standard Rescorla–Wagner learning model describes an agent's tendency to update prediction errors (PEs), the difference between feedback (F) and the prediction (P) of the agent on a certain trial (t).

$$\mathrm{PE}_t = F_t - P_t$$

The model that best described how participants learned about other persons in our previous studies[21,22], expanded the Rescorla–Wagner rule with pre-existing knowledge about the peer group. This prior knowledge was formalized as considering two sources of information: the *Reference Point* and *Granularity* of knowledge (described in more detail below). We surmise that the same information sources may be relevant while learning the word meanings (i.e., identifying the feature in question).

Granularity

Granularity refers to the level of detail with which participants represent previous knowledge. Coarse granularity assumes that a person applies the PE during learning to all items that fall within a (sub)category (e.g., all fast-food items). Fine granularity assumes that a person applies the PE to each individual item, but that the magnitude of this update depends on how similar the items are to the one they have received feedback about.

The similarity (SIM) between two items (*i*,*j*) is derived based on the item-level dissimilarity values (DIS) obtained from the MAT in the pre-experiment:

$$\mathrm{SIM}_{ij} = 1 - \mathrm{DIS}_{ij}$$

SIM was rescaled to the range of 0 (maximally dissimilar) and 1 (maximally similar). A depiction of SIM has been included in Fig. 1B right panel.

Reference points

Reference points refer to *a priori* expectations of an object's rating. This means that a person uses aggregated prior knowledge to infer what the non-word means. This *a priori* expectation may correspond to the true feature rating by either themselves (e.g., how colorful they think that the object is) or a representation of how people may rate this feature on average (e.g., how colorful do people think that the object is on average).

Here, we tested whether individuals relied more on self-ratings or population averages during learning. To this end, we investigated participants' own ratings of these features (i.e., self-ratings) and the mean feature values for each object from the pre-experiment (i.e., mean ratings) as potential reference points. A detailed description of our model fitting and model comparison approach, as well as model and parameter recovery, is included in the supplement.

Results

Results of the pre-experiment

The principal components analysis PCA yielded 69 components that cumulatively explained 90% of the variance of the original 120-item space. Components 1 and 2 reflected the distinction between the three main categories (i.e., activities, fashion, and food; Fig. 1C). Component 1 corresponded to a food-non-food contrast, while component 2 differentiated between activities and fashion items. The first two components or dimensions accounted for 19% of the total variance. The third dimension contrasted stationary, fitness equipment, and games against music and sound accessories (e.g., headphones); the fourth dimension represented the distinction between a healthier versus less healthy lifestyle; the fifth dimension contrasted an active outdoor lifestyle and fast food with less active, indoor activities and healthy foods. Descriptions and illustrations of components 3–10 extracted by the PCA can be found in Fig. S1. Overall, the PCA shed light on the rich semantic structure between objects based on the MAT task that extends beyond category and subcategory information.

A priori assigned categories and subcategories significantly predicted participants' arrangements of objects in the arena (item-pair dissimilarities predicted by category: adjusted $R^2 = 0.467$, $F(1,14398) = 12{,}630$, $p < 0.001$); standardized betas $(st.\beta) = -0.68$, 95% CI [−0.70, −0.67]; and by

subcategory *r*: adjusted $R^2 = 0.259$, $F(1,14398) = 5{,}028$, $p < 0.001$), $st.\beta = -0.51$, 95% CI = [−0.52, −0.49]). The category and subcategory information could be recovered from the reconstructed data based on the first *two* components (category: $\beta = 5.00$, SE = 0.083, $p < 0.001$, odds ratio = 16.99, 95% CI = [15.51, 18.65]; $R^2_{McF} = 0.562$, $\chi^2(2) = 3883.6$, $p < 0.001$; subcategory: $\beta = 6.14$, SE = 0.28, $p < 0.001$, odds ratio = 17.04, 95% CI = [13.92, 21.24]; $R^2_{McF} = 0.394$, $\chi^2(2) = 1659.7$, $p < 0.001$), but these components only explained 19% of the total variance in participants arrangements. The pre-experiment thus revealed a fine-grained semantic relatedness structure that cannot be sufficiently captured by the *a priori* defined coarse semantic structure (i.e., category and subcategory assignments). In the main experiment, we directly tested whether individuals relied on coarse category information or on fine-grained representation of object-level relationships during implicit learning.

Results of the main experiment

Changes in task-based prediction errors over time. Participants successfully showed learning for word meanings, evidenced by a significant reduction in their PEs over time and a negative correlation between PEs and trial number (colorful: Pearson's $r = -0.61$, CI = [−0.77, −0.38], $p < 0.001$; large: Pearson's $r = -0.84$, CI = [−0.91, −0.73]; $p < 0.001$; see Fig. 2B). Word clouds depicted the participants' notion of the concept in question, which was queried in an open answer question after each word run. Remarkably, some participants were able to explicitly label the feature in question (21% of participants correctly labeled the feature "colorful", and 6% correctly reported "large" as the feature in question).

Learning induced shifts in object similarity representations

In model 1, we investigated whether the learning task influenced post-learning item-dissimilarity representations by predicting item-pair dissimilarities post-learning based on category and, additionally, item-by-item concept dissimilarity. Replicating our findings from the pre-experiment, the category significantly predicted the proximity of item dissimilarity representations. Items of the same category (i.e., activity and foods) were placed closer together (adjusted $R^2 = 0.743$, $F(7,1714) = 711.7$, $p < 0.001$); standardized beta ($st.\beta$) = 1.67, $p < 0.001$, 95% CI = [1.62, 1.72], $\eta_p^2 = 0.716$). Moreover, item-level dissimilarity of features (colorful and large) additionally predicted item placement in the MAT (color: $st.\beta = 0.12$, $p < 0.001$, 95% CI = [0.09, 0.15], $\eta_p^2 = 0.005$; large: $st.\beta = 0.076$, SE = 0.018, $p < 0.001$, *std.* $\beta = 0.18$, $p < 0.001$, 95% CI = [0.14, 0.22], $\eta_p^2 = 0.014$), indicating that objects were placed further apart if they differed in the level of colorfulness and largeness. A significant interaction between category and feature dissimilarity meant that objects within the same category were additionally sorted by how colorful and large they were (color: *st.* $\beta = -0.14$, $p < 0.001$, 95% CI = [−0.19, −0.09], $\eta_p^2 = 0.005$; large: $st.\beta = -0.12$, $p < 0.001$, 95% CI = [−0.17, −0.07], $\eta_p^2 = 0.004$; Fig. 2C).

Model 2 (adjusted $R^2 = 0.645$, $F(4,1714) = 447.7$, $p < 0.001$) tested whether feature dissimilarity influenced item (dis)similarity judgements in the main experiment (post learning) more than in the pre-experiment (in the absence of learning). To this end, we set up a GLM, which predicted item-pair dissimilarities in the MAT of the main experiment with those from the pre-experiment and additionally with category and feature dissimilarity information (item-pair dissimilarity of colorfulness and largeness). The pre-experiment arrangements significantly predicted those in the main experiment ($\beta = 0.52$, $p < 0.001$, *st.* $\beta = 0.30$, 95% CI = [0.20, 0.40], $\eta_p^2 = 0.624$), confirming that object arrangements were robust across experiments. A significant three-way interaction between the pre-experiment arrangements, category, and feature dissimilarity ($\beta = -0.14$, $p = 0.04$, *st.* $\beta = -0.11$, 95% CI = [−0.21, 0] $\eta_p^2 = 0.001$) indicated that for item-pairs of the same category, feature dissimilarity exerted additional influence on item placements. This means that the greater the difference between item-pair dissimilarities in the pre-experiment was, the greater the influence of feature dissimilarity on item placements in the main experiment. To provide more detail on the interaction effect between pre- and main experiments, we tested the differential effects of the features used in the learning tasks on item placements in both the pre- and main experiments (see supplement). In summary, these results show that feature similarity is only applied in the main experiment in the expected direction. when items are semantically unrelated. Only in the main experiment is feature similarity used in the expected direction—the greater the feature similarity, the closer items are placed together (see Fig. S4 in the supplement).

Computational modelling results

We tested five models and additional variations. The models and parameters were recoverable. Please refer to the supplement for a detailed description of the main models and their additions. Bayesian model comparisons revealed Model 4 [Fine Granularity] as the best fitting model according to the random-effects analyses (see Fig. 3A). While Model 5 [Fine Granularity and Self Reference Point] was the best model using the fixed effects model comparison method, the random effects comparison takes the frequency of a model providing the best fit for participants' data into account. Model 4 provided the best model fit for participants' data compared to Model 5 (see Fig. 3A). The best fitting model, Fine Granularity, assumes no reliance on previous knowledge about the object features probed in this task. The model, however, scales the extent to which participants updated their estimates based on feedback by the fine-grained similarity between items in the set. In line with our previous findings on social learning, these results indicate that participants used a representation of item-level similarities during feature identification learning[21,22]. Importantly, participants' strategies deviated from the prescribed strategy, which is to rely on colorful and large estimates directly. The difference between the best-performing model in the set and the one that best describes participants' behavior stems from participants' implicitly learning about the feature in question. Participants were initially asked to rate how much a nonword applied to objects without being told the meaning of the nonword. Through task feedback, they could map the feature in question to the nonword they were asked to learn about (see supplement for more details).

Discussion

Using a multi-dimensional scaling approach, we discovered that the spatial representation of object similarity captured both broader predefined semantic categories and more nuanced semantic dimensions, such as one associated with a healthy, active lifestyle. Even finer-grained item-level similarity maps played a significant role in implicitly learning about specific object features. Interestingly, the features introduced in the learning task were recovered from participants' post-learning similarity arrangements, showing how knowledge representations can be refined through learning.

Spatial similarity representations reflect semantic relationships

Semantic relationships, also knowledge structures or cognitive schemata, play a crucial role in human experiences across various cognitive domains[12,42,43]. In our study, we observed that the most prominent semantic dimensions aligned with our predefined object categories and subcategories. Notably, further dimensions represented holistic aspects of various lifestyles, unveiling intricate connections between objects. We found a distinction between a less active lifestyle encompassing stationery, music, board games, junk food, and desserts, versus a fitness-oriented and healthy meals dimension. Our findings align with previous studies, which highlight the ability of multi-dimensional scaling tasks to reveal rich semantic structure that ties in participants' personal experiences, which are deeply rooted in social context[24,25,27,44]. The existence of spatial "cognitive maps" which efficiently organize knowledge is well-documented in the literature[45–49]. These maps guide attention and learning across different domains[45,46,50]. While two-dimensional representations are often emphasized in cognitive tasks, research suggests that cognitive representations are multi-dimensional and can be compressed or unfolded based on task demands[17,51]. Our study confirmed the utility of spatial representations in organizing cognitive structures. Subsequently, we investigated how this spatial representation is utilized during the learning process.

Object-similarity representations guide implicit feature identification learning

The current study directly shows that previous knowledge, also coined cognitive schemata, plays a crucial role in actively learning about object features. Our study, therefore, extends previous research that showed the relevance of knowledge structures or cognitive schemata in visual perception[16,52,53], reward learning[54,55], and the social domain[8,56]. Specifically, this study extended the computational modeling framework for social learning[21,22] to non-social, semantic knowledge.

Individuals relied on a fine-grained Similarity Model to learn word meanings through trial-by-trial feedback. Learning was driven by the prediction error (PE signal, i.e., the learning signal in reinforcement learning models across social and non-social domains[56,57]. Participants also represented fine-grained knowledge about object similarity during learning and utilized this knowledge to scale PE updating. This highlights the role of semantic relationship structure in inferring word meaning. Our findings align with the language learning literature, which has demonstrated that knowing the meaning of a word through a word-to-concept mapping does not replace the necessity of representing semantic knowledge[58,59]. Semantic relatedness, measured by word co-occurrence patterns in text corpora, can be learned through experience via unsupervised statistical learning[60,61] and plays an important role in language acquisition and reading comprehension[60,62].

Notably, participants' learning strategy deviated from the best-performing model in the set, which was the simple mapping of the feature in question to the respective non-word. The fact that participants did not show this direct mapping of feature to non-word means that participants did not know the meaning of the non-word from the beginning and instead learned by integrating PE signals into future inferences. Significant PE reductions over the course of the task and the open-answer question post-learning corroborated that participants indeed learned the meanings of the non-words throughout the task. Some participants were able to correctly report on the feature in question after learning. This verbal report is evidence for an explicit representation of the feature in question. While building such an explicit representation by inferring the actual feature in question is a desirable task outcome, we conjecture that reporting a different feature is not, per se, evidence that participants did not learn the task. Given that overall participants significantly reduced PEs over time, and a body of literature that describes implicit and explicit learning and mental representations as largely independent[1,63–65], we take significant reductions in prediction errors as evidence for successful implicit learning, which, in line with this literature, may not necessarily transfer into detailed explicit representations of the feature in question.

Object-similarity maps serve as flexible knowledge frameworks

In this study, we demonstrate that object similarity, i.e., semantic relatedness, is utilized during learning, and the learned features are reflected in item-level arrangements after learning. Object similarity maps are based on object co-occurrence[16]. In statistical learning, objects that occur most frequently together are those that share semantic features with each other. This co-occurrence information is deeply ingrained in the visual system[16]. Object similarity representations scaffold knowledge to aid learning of new information, as well as aiding associative inference. Associative inference refers to the ability to derive new information by forming links between known and related information. For example, if two stimuli share one property, it can be inferred that they also share another property[11,66].

Highlighting the importance of object co-occurrence and semantic relationships, neurons within the visual system, including in object- and scene-selective visual cortices, are tuned to the natural statistics of object contexts and frequencies[67]. The spatial representation of knowledge is supported by both the hippocampus and the medial prefrontal cortex (MPFC). Recent studies on the neural encoding of knowledge representations found that the hippocampus and the MPFC represent the underlying dimensions that organize complex abstract stimuli[17,68].

Flexible updating of the object-similarity structure observed in our study may depend on hippocampal–prefrontal integration. Recent theoretical models[69,70] and empirical work[17,70] suggest that the hippocampus encodes specific events and represents relationships between events at various levels of detail. This is in line with the logic of the Similarity learning framework that we applied in this study. Similarity models employ representations of coarser and fine-grained knowledge dimensions during learning (see ref. 51).

The short learning blocks induced a shift in individuals' object similarity representation post-learning. This is in line with previous findings showing that recurrent connections between the hippocampus and the prefrontal cortex facilitate access and updating of already acquired knowledge[11,71]. As new events are encoded, prior memories are re-encoded and reshaped by current events[72]. The current study probed whether

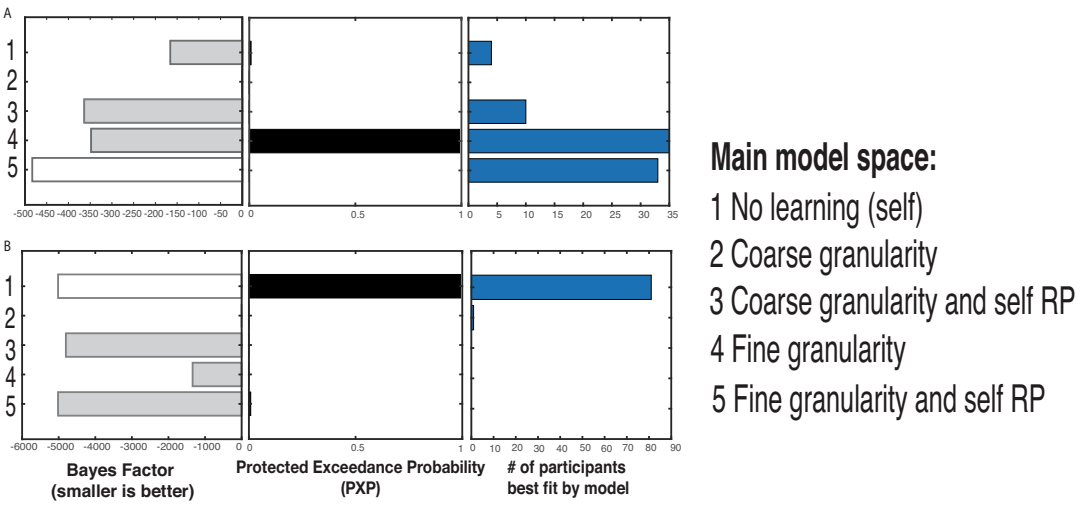


Fig. 3 | Bayesian model comparisons of the five main models' performance. The worst-performing model is used as the baseline. Smaller Bayes factors (BF) indicate a better fit. Left panel: Model comparison using fixed-effects analysis as the comparison metric. Middle panel: Random effect model comparison via posterior exceedance probability of the models. Right panel: the frequency of each model being the best model. **A** Model comparison on participants' data ($N = 82$ participants) using fixed-effects and random effects reveals that Model 5 [Fine granularity] performs best. This learning model uses fine-granularity information (item-similarities) to generalize across objects. **B** Best performing model or strategy to perform the task. Given that participants were learning about the feature ratings, the simple no-learning model that applies the feature rating to a specific item in question was the best-performing strategy.

learning to deduce abstract features in question relied on the same cognitive mechanisms as social learning—such as inferring others' preferences or traits from feedback[21,22,37]. In these prior tasks, participants learned about different people or task profiles and various items. Indeed, we found that, similar to learning in social contexts, participants relied on a reinforcement learning model that weighed feedback based on item similarity. From this, we can conclude that the learning mechanism generalizes across stimulus sets (trait-, preferences, non-social features), and social versus nonsocial contexts. Another important next step will also be to corroborate the key regions involved in the flexible deployment and updating of object-similarity structures during learning.

Limitations

We have assessed semantic relationships with the MAT task, which has been extensively used and validated against other measures of semantic similarity such as pairwise judgements (see ref. 73 e.g.,). However, all measures of semantic similarity have an important caveat –they are context dependent, meaning that the similarity ratings depend on the specific item set in question[31]. It is therefore important to test the generalization of our results across various tasks and item sets to establish more accurate concept similarity measures.

Due to the short learning intervention, there was a small effect of the learned concept on pre-existing knowledge representation. While we show that the learning modeled in the RL framework produced a shift in object similarity representation post-learning, we cannot conclusively disambiguate the roles of attentional shifts to a feature in question during the learning task (i.e., priming) and trial-by-trial learning as the primary contributor to the observed shifts in similarity representations. Future studies should corroborate these learning-induced shifts in more extensive learning interventions. An intriguing possibility is that the reorganization of item similarity structures through learning or priming represents a general mechanism for increasing our prior knowledge and behavioral adaptation across task contexts[51,74–77]).

Conclusion

In conclusion, this study elucidates how humans organize semantic knowledge. Participants' spatial representations of semantic relatedness, i.e., their semantic relatedness map, guided implicit learning of object features. Moreover, item-level relationship structures were flexibly rearranged through learning, possibly reflecting a broader mechanism of how knowledge is activated for learning and flexibly reorganized as a function of the learning content.

Statement of relevance

Humans possess the remarkable ability to retain both generalized knowledge about categories of items (e.g., apples and bananas are fruits) and detailed item-specific similarities (e.g., peaches resemble apples more than bananas). Generalizations or schemata and fine-grained knowledge for specifics profoundly impact our decision-making and learning. Schemata, as mental shortcuts, expedite decisions, thereby enhancing efficiency. Simultaneously, preserving detailed item-level similarities enables us to efficiently zoom in on specific features in question. The current study reveals the interplay between the structure of semantic knowledge and learning. We show that individuals represent item-level similarity information during implicit feature identification learning and that learning dynamically updates the existing item similarity representations. These results carry far-reaching implications for how humans access and build knowledge across a variety of cognitive and social domains. By adapting and fine-tuning representations of object similarities through learning, we continuously refine cognitive frameworks, enabling more effective decision-making and knowledge integration. In short, the current study establishes that semantic relatedness is used in implicit learning and that learning, in turn, produces shifts in semantic relatedness representations.

Data availability

Data can be found on the lab's GitHub account (https://doi.org/10.17605/OSF.IO/6NWQU).

Code availability

Analysis scripts can be found on the lab's GitHub account (https://doi.org/10.17605/OSF.IO/6NWQU).

Received: 29 July 2024; Accepted: 30 April 2025;
Published online: 13 May 2025

References

1. Berry, D. C. & Broadbent, D. E. On the Relationship between Task Performance and Associated Verbalizable Knowledge. *Q. J. Exp. Psychol. Sect. A* **36**, 209–231 (1984).
2. Lynn, C. W., Kahn, A. E., Nyema, N. & Bassett, D. S. Abstract representations of events arise from mental errors in learning and memory. *Nat. Commun.* **11**, 2313 (2020).
3. Nah, J. C. & Shomstein, S. Target frequency modulates object-based attention. *Psychon. Bull. Rev.* **27**, 981–989 (2020).
4. Reber, A. S. More thoughts on the unconscious: reply to Brody and to Lewicki and Hill. *J. Exp. Psychol. Gen.* **118**, 242–244 (1989).
5. Seger, C. A. & Seger, C. A. Efficiency and conceptual fluency as independent mechanisms in implicit learning. *Dissertation Abstr. Int.: Sect. B: Sci. Eng.* **55**, 5107 (1995).
6. Glaze, C. M., Kable, J. W. & Gold, J. I. Normative evidence accumulation in unpredictable environments. *Elife* **4**, e08825 (2015).
7. Seger, C. A. Implicit learning. *Psychol. Bull.* **115**, 163–196 (1994).
8. FeldmanHall, O. et al. Stimulus generalization as a mechanism for learning to trust. *Proc. Natl Acad. Sci. USA* **115**, E1690–E1697 (2018).
9. Franklin, N. T., Norman, K. A., Ranganath, C., Zacks, J. M. & Gershman, S. J. Structured event memory: a neuro-symbolic model of event cognition. *Psychol. Rev.* **127**, 327–361 (2020).
10. Kronenfeld, D. B., Schank, R. C. & Abelson, R. P. Scripts, plans, goals, and understanding: an inquiry into human knowledge structures. *Language* **54**, 779 (1978).
11. Bowman, C. R., Iwashita, T. & Zeithamova, D. Tracking prototype and exemplar representations in the brain across learning. *Elife* **9**, e59360 (2020).
12. Mayer, J. D. & Bower, G. H. Learning and memory for personality prototypes. *J. Pers. Soc. Psychol.* **51**, 473–492 (1986).
13. Nah, J. C., Malcolm, G. L. & Shomstein, S. Task-irrelevant semantic properties of objects impinge on sensory representations within the early visual cortex. *Cereb. Cortex Commun.* **2**, tgab049 (2021).
14. Wegner-Clemens, K., Malcolm, G. & Shomstein, S. Search efficiency scales with audiovisual semantic relatedness in a continuous manner. *Psychophysics* **79**, 154 (2024).
15. Frank, D., Montaldi, D., Wittmann, B. & Talmi, D. Beneficial and detrimental effects of schema incongruence on memory for contextual events. *Learn. Mem.* **25**, 352–360 (2018).
16. Bonner, M. F. & Epstein, R. A. Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nat. Commun.* **12**, 4081 (2021).
17. Garvert, M. M., Saanum, T., Schulz, E., Schuck, N. W. & Doeller, C. F. Hippocampal spatio-predictive cognitive maps adaptively guide reward generalization. *Nat. Neurosci.* **26**, 615–626 (2023).
18. Kahnt, T. & Tobler, P. N. Dopamine regulates stimulus generalization in the human hippocampus. *Elife* **5**, e12678 (2016).
19. Arana, S., Hagoort, P., Schoffelen, J.-M. & Rabovsky, M. Perceived similarity as a window into representations of integrated sentence meaning. *Behav. Res. Methods* https://doi.org/10.3758/s13428-023-02129-x (2023).
20. Lenci, A. Distributional models of word meaning. *Annu. Rev. Linguist.* **4**, 151–171 (2018).

21. Frolichs, K. M. M., Rosenblau, G. & Korn, C. W. Incorporating social knowledge structures into computational models. *Nat. Commun.* **13**, 6205 (2022).
22. Rosenblau, G., O'Connell, G., Heekeren, H. R. & Dziobek, I. Neurobiological mechanisms of social cognition treatment in high-functioning adults with autism spectrum disorder. *Psychol. Med.* **50**, 2374–2384 (2020).
23. Wise, T., Charpentier, C. J., Dayan, P. & Mobbs, D. Interactive cognitive maps support flexible behavior under threat. *Cell Rep.* **42**, 113008 (2023).
24. Kriegeskorte, N. Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
25. Parkinson, C., Kleinbaum, A. M. & Wheatley, T. Spontaneous neural encoding of social network position. *Nat. Hum. Behav.* **1**, 1–7 (2017).
26. Popal, H., Wang, Y. & Olson, I. R. A guide to representational similarity analysis for social neuroscience. *Soc. Cogn. Affect. Neurosci.* **14**, 1243–1253 (2019).
27. Tamir, D. I., Thornton, M. A., Contreras, J. M. & Mitchell, J. P. Neural evidence that three dimensions organize mental state representation: rationality, social impact, and valence. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.1511905112 (2016).
28. Thornton, M. A., Weaverdyck, M. E. & Tamir, D. I. The brain represents people as the mental states they habitually experience. *Nat. Commun.* **10**, 2291 (2019).
29. Nosofsky, R. M. Attention, similarity, and the identification–categorization relationship. *J. Exp. Psychol. Gen.* **115**, 39–57 (1986).
30. Goldstone, R. L. Effects of categorization on color perception. *Psychol. Sci.* **6**, 298–304 (1995).
31. Tversky, A. Features of similarity. *Psychol. Rev.* **84**, 327–352 (1977).
32. Dayan, P. & Niv, Y. Reinforcement learning: the good, the bad and the ugly. *Curr. Opin. Neurobiol.* **18**, 185–196 (2008).
33. Lockwood, P. L. & Klein-Flügge, M. Computational modelling of social cognition and behaviour—a reinforcement learning primer. *Soc. Cogn. Affect Neurosci.* **16**, 761–771 (2021).
34. Ruff, C. C. & Fehr, E. The neurobiology of rewards and values in social decision making. *Nat. Rev. Neurosci.* **15**, 549–562 (2014).
35. Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J. P. & Lamm, C. Using reinforcement learning models in social neuroscience: frameworks, pitfalls and suggestions of best practices. *Soc. Cogn. Affect Neurosci.* **15**, 695–707 (2020).
36. Dayan, P. & Berridge, K. C. Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cogn. Affect. Behav. Neurosci.* **14**, 473–492 (2014).
37. Rosenblau, G., Korn, C. W. & Pelphrey, K. A. A computational account of optimizing social predictions reveals that adolescents are conservative learners in social contexts. *J. Neurosci.* **38**, 974–988 (2018).
38. Kriegeskorte, N. & Mur, M. Inverse MDS: inferring dissimilarity structure from multiple item arrangements. *Front. Psychol.* **3**, 245 (2012).
39. Mur, M. et al. Human object-similarity judgments reflect and transcend the primate-IT object representation. *Front. Psychol.* **4**, 128 (2013).
40. Lee, W. & Grimm, K. J. Generalized linear mixed-effects modeling programs in R for binary outcomes. *Struct. Equ. Model.* **25**, 824–828 (2018).
41. Michas, I. C. & Berry, D. C. Implicit and explicit processes in a second-language learning task. *Eur. J. Cogn. Psychol.* **6**, 357–381 (1994).
42. Cantor, N. & Mischel, W. Traits as prototypes: effects on recognition memory. *J. Pers. Soc. Psychol.* **35**, 38–48 (1977).
43. Fiske, S. T. & Linville, P. W. What does the schema concept buy us? *Pers. Soc. Psychol. Bull.* **6**, 543–557 (1980).
44. Joshanloo, M. The structure of the MHC-SF in a large American sample: contributions of multidimensional scaling. *J. Ment. Health* **29**, 139–143 (2020).
45. Bellmund, J. L. S., Gärdenfors, P., Moser, E. I. & Doeller, C. F. Navigating cognition: spatial codes for human thinking. *Science (1979)* **362**, eaat6766 (2018).
46. Constantinescu, A. O., O'Reilly, J. X. & Behrens, T. E. J. Organizing conceptual knowledge in humans with a gridlike code. *Science (1979)* **352**, 1464–1468 (2016).
47. Tavares, R. M. et al. A map for social navigation in the human brain. *Neuron* **87**, 231–243 (2015).
48. Thornton, M. A., Rmus, M., Vyas, A. D. & Tamir, D. I. Transition dynamics shape mental state concepts. *J. Exp. Psychol. Gen.* **152**, 2804–2829 (2023).
49. Wu, C. M., Schulz, E., Garvert, M. M., Meder, B. & Schuck, N. W. Similarities and differences in spatial and non-spatial cognitive maps. *PLoS Comput. Biol.* **16**, e1008149 (2020).
50. Pettine, W. W., Raman, D. V., Redish, A. D. & Murray, J. D. Human generalization of internal representations through prototype learning with goal-directed attention. *Nat. Hum. Behav.* **7**, 442–463 (2023).
51. Rosenblau, G., Frolichs, K. & Korn, C. W. A neuro-computational social learning framework to facilitate transdiagnostic classification and treatment across psychiatric disorders. *Neurosci. Biobehav. Rev.* **149**, 105181 (2023).
52. Sadeghi, Z., McClelland, J. L. & Hoffman, P. You shall know an object by the company it keeps: an investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia* **76**, 52–61 (2015).
53. Stansbury, D. E., Naselaris, T. & Gallant, J. L. Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron* **79**, 1025–1034 (2013).
54. Palminteri, S., Khamassi, M., Joffily, M. & Coricelli, G. Contextual modulation of value signals in reward and punishment learning. *Nat. Commun.* **6**, 8096 (2015).
55. Chien, S., Wiehler, A., Spezio, M. & Gläscher, J. Congruence of inherent and acquired values facilitates reward-based decision-making. *J. Neurosci.* **36**, 5003–5012 (2016).
56. Joiner, J., Piva, M., Turrin, C. & Chang, S. W. C. Social learning through prediction error in the brain. *NPJ Sci. Learn.* **2**, 8 (2017).
57. Hampton, A. N., Bossaerts, P. & O'Doherty, J. P. Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proc. Natl Acad. Sci. USA* **105**, 6741–6746 (2008).
58. Clerkin, E. M. & Smith, L. B. Real-world statistics at two timescales and a mechanism for infant learning of object names. *Proc. Natl Acad. Sci. USA* **119**, e2123239119 (2022).
59. Morfidi, E., Mikropoulos, A. & Rogdaki, A. Using concept mapping to improve poor readers' understanding of expository text. *Educ. Inf. Technol. (Dordr)* **23**, 271–286 (2018).
60. Romberg, A. R. & Saffran, J. R. Statistical learning and language acquisition. *WIREs Cogn. Sci.* **1**, 906–914 (2010).
61. Sherman, B. E., Graves, K. N. & Turk-Browne, N. B. The prevalence and importance of statistical learning in human cognition and behavior. *Curr. Opin. Behav. Sci*. https://doi.org/10.1016/j.cobeha.2020.01.015 (2020).
62. Ricketts, J., Davies, R., Masterson, J., Stuart, M. & Duff, F. J. Evidence for semantic involvement in regular and exception word reading in emergent readers of English. *J. Exp. Child Psychol.* **150**, 330–345 (2016).
63. Apperly, I. A. & Butterfill, S. A. Do humans have two systems to track beliefs and belief-like states? *Psychol. Rev.* **116**, 953–970 (2009).
64. Willingham, D. B. & Goedert-Eschmann, K. The relation between implicit and explicit learning: evidence for parallel development. *Psychol. Sci.* **10**, 531–534 (1999).
65. Willingham, D. B., Nissen, M. J. & Bullemer, P. On the development of procedural knowledge. *J. Exp. Psychol. Learn. Mem. Cogn.* **15**, 1047–1060 (1989).
66. Zeithamova, D. & Bowman, C. R. Generalization and the hippocampus: more than one story? *Neurobiol. Learn. Mem.* **175**, 107317 (2020).

67. Chan, A. W.-Y., Kravitz, D. J., Truong, S., Arizpe, J. & Baker, C. I. Cortical representations of bodies and faces are strongest in commonly experienced configurations. *Nat. Neurosci.* **13**, 417–418 (2010).
68. Behrens, T. E. J. et al. What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* **100**, 490–509 (2018).
69. Mack, M. L., Love, B. C. & Preston, A. R. Building concepts one episode at a time: the hippocampus and concept formation. *Neurosci. Lett.* **680**, 31–38 (2018).
70. Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M. & Norman, K. A. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philos. Trans. R. Soc. B* https://doi.org/10.1098/rstb.2016.0049 (2017).
71. Koster, R. et al. Big-loop recurrence within the hippocampal system supports integration of information across episodes. *Neuron* **99**, 1342–1354.e6 (2018).
72. Hintzman, D. L. Judgment of frequency versus recognition confidence: repetition and recursive reminding. *Mem. Cogn.* **32**, 336–350 (2004).
73. Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J. F. & Charest, I. The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *Neuroimage* **194**, 12–24 (2019).
74. Goldstone, R. L. Influences of categorization on perceptual discrimination. *J. Exp. Psychol. Gen.* **123**, 178–200 (1994).
75. Thornton, M. A. & Tamir, D. I. Mental models accurately predict emotion transitions. *Proc. Natl Acad. Sci. USA* **114**, 5982–5987 (2017).
76. Tamir, D. I. & Thornton, M. A. Modeling the predictive social mind. *Trends Cogn. Sci.* **22**, 201–212 (2018).
77. Thornton, M. A. & Tamir, D. I. People represent mental states in terms of rationality, social impact, and valence: Validating the 3 d Mind Model. *Cortex* **125**, 44–59 (2020).

Acknowledgements

This study was conducted in the context of a larger project on modeling social and non-social learning in autism, which is funded by the National Institute for Mental Health (NIMH, R01MH116252). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the paper.

Authors contribution

J.K.D. managed and set up online experiments, led the statistical analysis and writing of the paper. S.S. oversaw the project progress and contributed to the interpretation of results and writing of the paper. The principal investigator, G. R., acquired the funding to support the project, oversaw the project progress, led the computational modeling analyses and writing of the paper.

Competing interests

We thank Jasper van den Bosch for his assistance with integrating the multi-arrangement task into our experimental pipeline. We also thank Yen-Wen Chen for her help with organizing the code for data sharing and Ella Chapman for her help with formatting and proofreading.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44271-025-00259-w.

Correspondence and requests for materials should be addressed to Jonathan K. Doyon or Gabriela Rosenblau.

Peer review information *Communications Psychology* thanks Guangyao Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Troby Ka-Yan Lui. [A peer review file is available.]