Task-relevant representational formats in multi-layered memory traces

Rebekka Heinen[1,*], Elias M.B. Rau[1,*] , Nora A. Herweg[1,*], Nikolai Axmacher[1,2]

1 Department of Neuropsychology, Institute of Cognitive Neuroscience, Faculty of Psychology, Ruhr University Bochum, 44801 Bochum, Germany.
2 State Key Laboratory of Cognitive Neuroscience and Learning and IDG/McGovern Institute for Brain Research, Beijing Normal University, 100875 Beijing, China.

* Equal contributions

Corresponding authors: Rebekka Heinen (rebekka.heinen@ruhr-uni-bochum.de), Nikolai Axmacher (nikolai.axmacher@ruhr-uni-bochum.de)

Keywords: memory traces, representational formats, deep neural networks, consolidation, targeted memory reactivation

ABSTRACT

During encoding, stimuli are embedded into memory traces that allow for their later retrieval. This process is selective, however, because not every aspect of our experiences can be remembered. In addition, post-encoding stages including consolidation are widely assumed to induce transformation processes of the memory trace. It is unclear, however, how selective the memory trace is, whether irrelevant information is completely removed during encoding and/or consolidation, and how this affects retrieval of either general (gist-like) or specific (perceptual) information. Here we show that memory traces consist of multi-layered representational spaces whose formats are flexibly strengthened or weakened during encoding and consolidation depending on task instructions, distinctly shaping their affordances for general or specific retrieval. In a series of behavioral experiments, participants first compared pairs of natural images on either two conceptual or two perceptual dimensions, leading them to incorporate the images into representational "spaces" defined by Euclidean distances. We found that distances in task-relevant but not irrelevant spaces affected memory strengths. Conceptual encoding benefitted general without impairing specific retrieval, suggesting that perceptual information remained in the memory trace even if it was task-irrelevant. By contrast, targeted memory reactivation (TMR) of conceptual encoding improved memory strength but deteriorated perceptual discrimination during retrieval, indicating that it weakened the accessibility of perceptual formats. Our results demonstrate the flexibility of representational formats that are incorporated into memory traces, and more generally show how the organization of information in representational spaces shapes human behavior.

Funding statement: R.H. is funded by the German Mercator Research Center Ruhr (MERCUR), project number Ex-2021-001; E.M.B.R. is funded by GIF I-1478-418.13/2018; N.A.H. received funding by ERC CoG 864164; N.A. acknowledges funding by ERC CoG 864164.

Supplementary material: All supplementary materials have been appended to this document and can be found after the reference section.

Data/code availability: All data and code relating to this paper can be found on the open science framework: https://osf.io/qn2gw/?view_only=2ca202d43ad8412e86282e020d78068f

Main

When you think back to your last birthday party, you may be able to mentally re-experience numerous details from this event – the taste of the cake, the melodies of the music, and the color of the wrapping paper – and these memories may already fuel imaginations of your next birthday. Episodic memories do not only help us remember past experiences but also provide a guide for future behavior. They rely on the reactivation of memory traces, which can be detected in the human brain via multivariate analysis methods such as representational similarity analysis (RSA)[1,2] and which guide behavioral memory decisions, yet not all information about an episode is stored: Remembering every detail would not only exceed the capacity of the brain, but also contradict the need to generalize past information to future situations. However, it is still not clear which information from an episode is maintained in memory. According to some theoretical frameworks, memory traces are "sparse" and only contain novel or unexpected information, which is complemented at the time of retrieval by further contents from pre-existing semantic knowledge[3,4]. Alternatively, memory traces may initially contain rich sensory information that degrades slowly over time[5,6]. Finally, the amount and kind of information that is stored in memory traces may vary depending on the behavioral goals during an experience[7]. Here we set out to test these competing hypotheses on the nature of the memory trace.

During perception, sensory information is hierarchically processed in the brain. For visual content, this hierarchy ranges from low-level perceptual information about e.g. edges and colors via more complex features such as textures or shapes to conceptual and semantic information. These features are extracted along the posterior-to-anterior extent of the ventral visual pathway (VVP) and lead to multiple coexisting representational formats of the same stimulus[5,8]. A similar processing hierarchy exists in convolutional deep neural networks (cDNN) trained for object recognition, providing a link between representations in the different cDNN layers and the various processing steps along the VVP[9–15]. In addition to these sensory formats of increasing complexity, semantic representations can be identified by corpus-based models of deep natural language processing (dNLP) that use text rather than images as input[16]. Thus, cDNNs and dNLPs provide different quantitative models of cognitive representations in a variety of representational formats[8,17–19]. Representational similarities in the individual formats correspond to distances in diverse representational "spaces" (with higher distances reflected by lower similarities) that can be employed for cognitive operations such as perception, memory, and reasoning[20].

While the neural basis and functional role of representational formats has been extensively studied during perception, fewer studies investigated the formats of memory traces, i.e., of representations in the absence of sensory inputs[21,22]. Crucially, certain representational

formats may be more relevant for subsequent tasks, and this may influence the availability and accessibility of information during later memory retrieval. Using DNN models to quantify perceptual and conceptual formats, it was recently shown that conceptual/semantic information is preferentially stored in a memory trace and supports later recall, while perceptual information is lost over time[18,23,24] (see Heinen et al., 2023 for review). This would suggest that memory traces are quickly transformed to only contain conceptual/semantic representational formats.

Prominent theoretical accounts have conceptualized these transformation processes as gist abstraction[25] or semantization[26,27]. These frameworks propose that transformation condenses detail-rich perceptual contents into more gist-like representations, because conceptual formats may be more generalizable and thus relevant for future behavior than perceptual formats. Processes of memory transformation and gist extraction have been linked to memory consolidation[28] which may start early after encoding[18,29] but is particularly pronounced during sleep[30]. Consolidation can be experimentally enhanced via targeted memory reactivation, i.e. by presenting auditory or olfactory cues during sleep that have been previously associated with stimuli[31,32]. However, are perceptual features necessarily lost in favor of conceptual features or do they remain silently available in the memory trace? In other words, is the representational format indeed subject to transformation (e.g., all sensory information is transformed into conceptual formats), or do multiple formats co-exist with some being gradually strengthened and others gradually weakened?

And is this process influenced by task demands? Indeed, various cognitive theories propose that representational formats depend on behavioral instructions during encoding. This has been suggested both in classical frameworks including levels of processing[33] and transfer-appropriate processing[34], but also in more recent theories based on reinforcement learning[35–37]. According to these accounts, task demands during encoding may flexibly determine which representational formats are embedded into a memory trace.

In addition, task demands during retrieval may influence which specific representational format is required: In some cases, it may be necessary to reactivate perceptually detailed information from a memory trace in order to distinguish previously experienced stimuli from perceptually similar new items (pattern separation or differentiation). Other situations may require generalized information about previously experienced categories, independent of the perceptual details of a particular exemplar (pattern completion or generalization). These diverging retrieval demands would benefit from richer memory traces with multiple different representational formats that can be flexibly reactivated depending on situational needs[23].

Here, we conducted a series of behavioral studies to address a fundamental question: how "shallow" or “deep” is a memory trace – i.e., does it contain only a single representational format or multiple formats, how are these formats determined, how may they support different retrieval demands, and (how) do they change after encoding? We investigated the representational formats during encoding, consolidation, and retrieval, and tested (1) whether behavioral goals during encoding affect the representational format of memory traces; (2) whether these formats serve different functional roles during retrieval; and (3) whether the formats that are incorporated into a memory trace during encoding can be subsequently modified via targeted memory reactivation.

Results

Selective embedding of memory traces into task-relevant representational spaces

In studies I and II, participants were presented pairs of natural images of animals and were asked to judge the similarity of these images on either two perceptual dimensions (perceptual blocks), on two conceptual dimensions (conceptual blocks), or on one perceptual and one conceptual dimension (mixed blocks, only in study I; Fig. 1b). During recognition, participants saw all images from the encoding phase together with an equal number of new items (Fig. 1c) and indicated their confidence that a stimulus was old or new on a visual analogue scale (Fig. 1c).

We first tested whether representational distances affected the ease of similarity judgments – i.e., whether responses to images were given faster when their distance on relevant dimensions were larger. Since we asked participants to perform similarity judgements on either two perceptual dimensions, two conceptual dimensions, or one perceptual and one conceptual dimension, we could compute the Euclidean distance of each image pair in its task-relevant 2D space (Fig. 1d). We used a linear mixed model to test the effect of task-relevant distance on reaction times including a random participant intercept. Significance testing was done using a likelihood ratio test between a full model, including the fixed effect of interest, and a reduced model without the fixed effect of interest (participant intercept only; see Methods for more details; beta weights of full linear models can be found in Supplement Table 1). Not surprisingly, we found that larger 2D distances of task-relevant dimensions predicted faster response times (Study I: $z = -11.28$, $\chi^2_{(1)} = 126.69$, $p < 0.001$; Study II: $z = -6.67$, $\chi^2_{(1)} = 44.49$, $p < 0.001$; Fig. 1e). Further, we tested whether behavior in mixed spaces would differ from their conceptual or perceptual space counterparts. We did not find a difference between judgements in purely conceptual or perceptual spaces with their respective dimension in mixed spaces (see Supplementary Fig. 1). We thus decided to use conceptual and perceptual spaces only for further analyses.

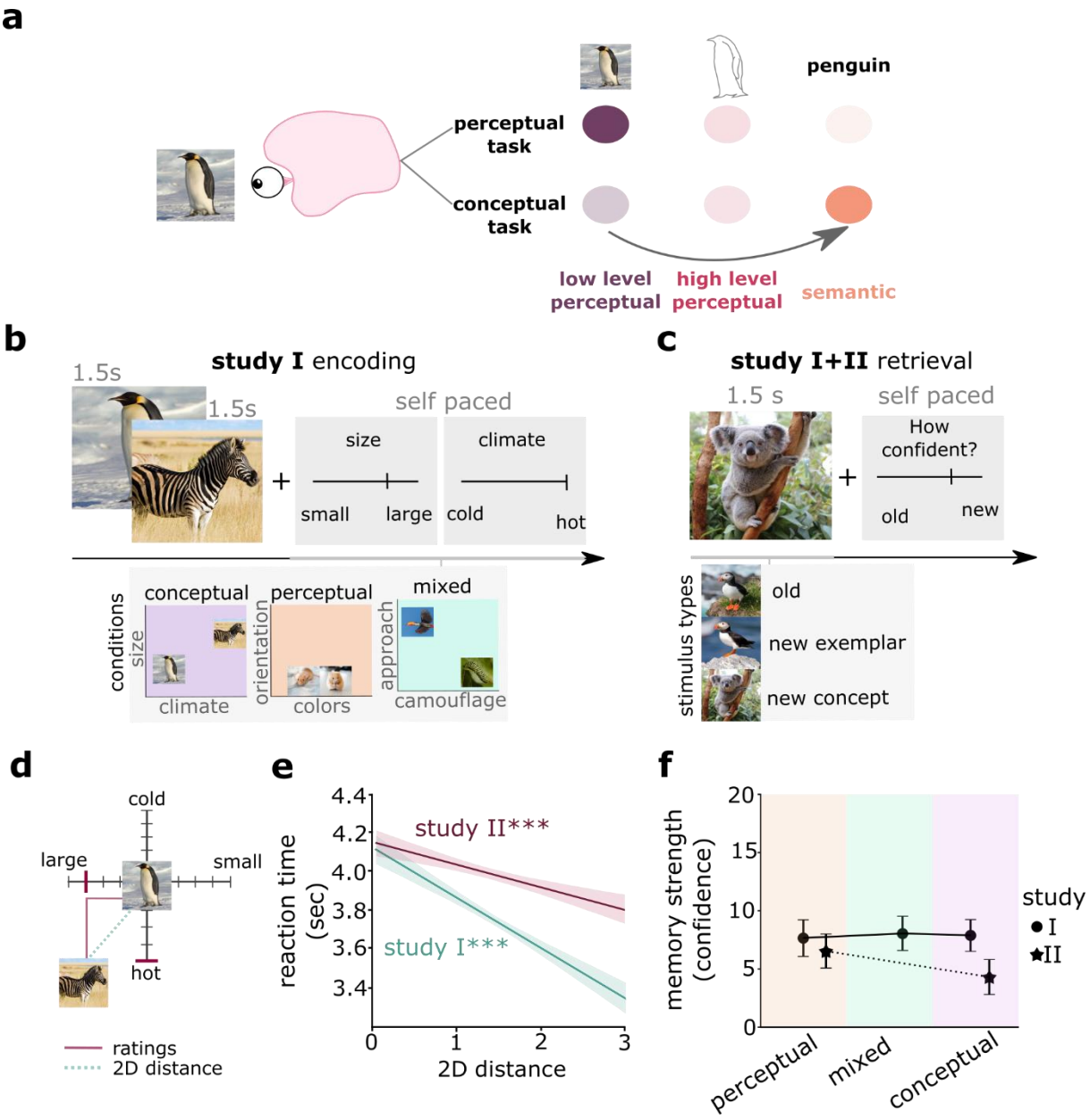


Figure 1: Embedding of task-relevant representational formats into memory traces. a) Theoretical framework: We hypothesize that memory traces consist of multiple representational formats whose accessibility can be flexibly adapted depending on task-demands. Perceptual tasks may increase the accessibility of perceptual features (left), while conceptual tasks may enhance processes of gist-abstraction and semantization (right). b) Encoding phase in studies I+II: Participants encoded natural images of animals in either perceptual or conceptual conditions, emphasizing different task-relevant representational formats. In each trial, participants rated the similarity of image pairs on 2 different perceptual and/or conceptual dimensions via visual analogue scales. Study I comprised blocks with 2 perceptual features, blocks with 2 conceptual features, and mixed blocks. Study II contained only perceptual and conceptual blocks, during which 2 different sounds were played for later cueing during sleep (see Figure 4 for more details). c) Recognition phase in studies I+II: Participants indicated their confidence whether an image was old or new on a visual analogue scale. d) Representational distances were derived from similarity judgments in each single dimension and from the Euclidean distance of both judgments (2D distance). e) Higher 2D distances are associated with shorter reaction times during encoding. f) Memory strength did not depend on encoding tasks. Data are visualized after removing participant-wise estimated random effects. 95% confidence intervals, error bars indicate SEM.***, p<0.001

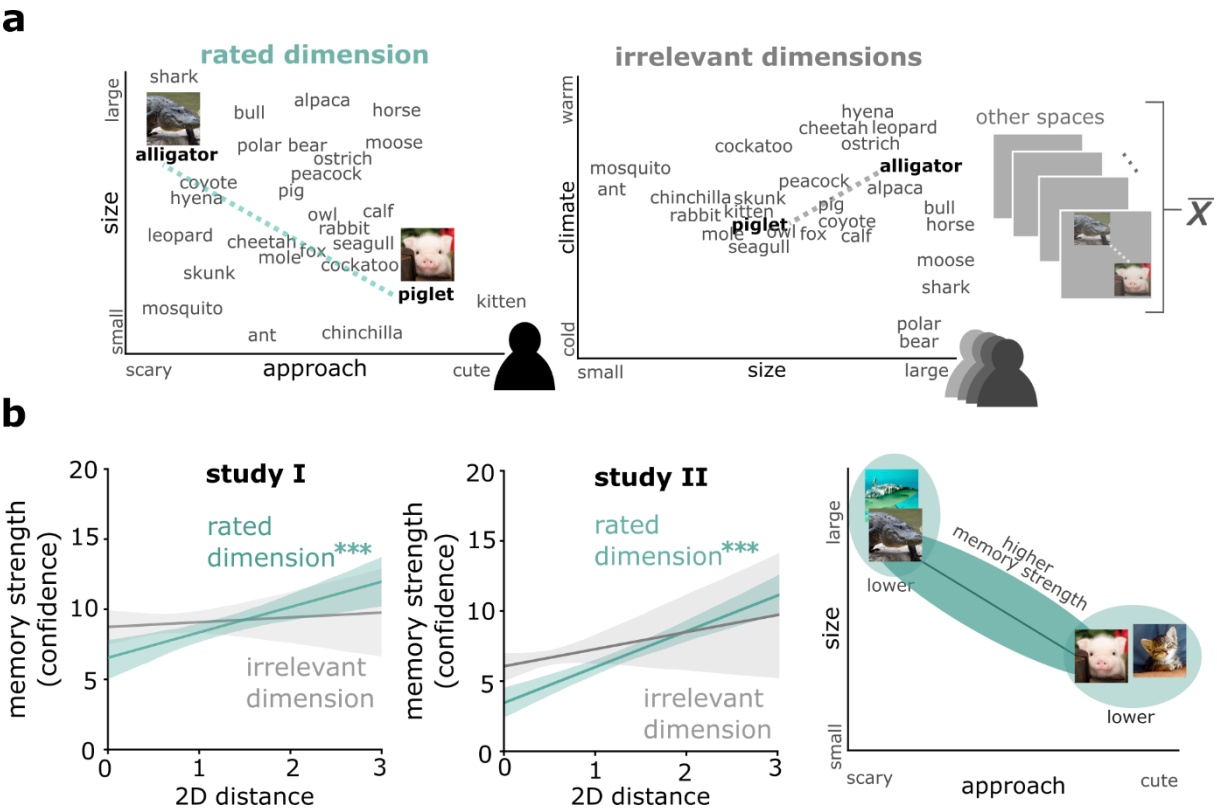


Figure 2: Embedding of task-relevant representational spaces into memory traces. a) Analysis of task-relevant and task-irrelevant encoding spaces. Left: 2D distances in task-relevant encoding spaces were derived from participant-specific similarity judgments (figure shows space in example participant). Right: 2D distances in task-irrelevant spaces were computed as average distance in least square solutions of all other encoding spaces in all other participants (figure shows one corresponding task-irrelevant space across all other participants). b) Memory strength depends on 2D distances in task-relevant but not task-irrelevant encoding spaces with better memory for higher distances. Left: Study I, Middle: Study 2, Right: schematic depiction. Data are visualized after removing participant-wise estimated random effects. 95% confidence intervals. ***, p<0.001

Next, we investigated whether these task-relevant spaces did not only explain the ease of behavioral ratings during encoding, but whether they were also selectively incorporated into memory traces – i.e., whether the accessibility of individual items for recognition memory judgements was influenced by their task-relevant encoding space. We first analyzed whether the encoding space of an item (perceptual or conceptual) affected the memory strength of that item, as the levels of processing account would predict superior memory for images encountered in conceptual vs. perceptual. Using a linear mixed model of trial-wise memory strength with “encoding task” as a predictor and participant-specific slopes as random effects, we did not find effects for more accessible memory traces following conceptual vs. perceptual judgements (Study I: $z = -0.36$, $\chi^2_{(1)} = 0.42$, $p = 0.516$; Study II: $z = 0.15$, $\chi^2_{(1)} = 0.02$, $p = 0.874$; likelihood ratio test; Fig. 1f). This indicates that conceptual encoding instructions per se did not influence the subsequent accessibility of memory traces, pointing to a possible role of other factors such as the task-relevance of a particular format during encoding.

Next, we analyzed whether memory traces of individual items involved the fine-grained structure of task-relevant representational spaces, by testing whether representational distances in the respective task-relevant format during encoding influenced subsequent recognition memory performance. For each pair of images, we extracted both the task-relevant distances (i.e., the participant-specific ratings of their similarity) and task-irrelevant distances, which were calculated as the average distance of the same pair of images from a least squares solution over all rated images in all other spaces not rated by a given participant (Fig. 2a). We then computed a linear mixed model to predict trial-wise memory strengths using the task-relevant distances (2D representational distances) as predictors and participant-specific slopes as random effects. We found that distances in task-relevant but not task-irrelevant formats predicted subsequent recognition memory strength (Fig. 2b). In both Study I+II, higher task-relevant representational distances predicted better recognition memory performance (Study I: $z = 4.16$, $\chi^2_{(1)} = 17.32$, $p < 0.001$; Study II: $z = 5.74$, $\chi^2_{(1)} = 33.02$, $p < 0.001$; likelihood ratio test). In study II, this effect was significant for both the first ($z = -4.65$, $\chi^2_{(1)} = 21.65$, $p < 0.001$) and the second image in a trial ($z = -4.46$, $\chi^2_{(1)} = 19.94$, $p < 0.001$); in study I, it was significant for the second image ($z = -4.27$, $\chi^2_{(1)} = 18.27$, $p < 0.001$) and by trend for the first image ($z = -1.64$, $\chi^2_{(1)} = 2.71$, $p = 0.099$). By contrast, when including the average distance of the same images in task-irrelevant representational spaces as predictors in the linear mixed model, we did not find a significant effect (Study I: $z = -0.89$, $\chi^2_{(1)} = 0.80$, $p = 0.373$; Study II: $z = 1.25$, $\chi^2_{(1)} = 1.58$, $p = 0.208$; Fig. 2b; see Supplementary Fig. 2a-b for separate results in the two encoding spaces). Thus, the irrelevant distances did not predict memory, independent of the processing format. While relevant distances predicted memory confidence for old images, we did not find an effect of relevant distance on correctly rejecting new exemplars as new (Study I: $z = -1.51$, $\chi^2_{(1)} = 2.31$, $p = 0.128$; Study II: $z = -0.78$, $\chi^2_{(1)} = 0.61$, $p = 0.433$).

These results show that task demands do not only affect similarity judgements during encoding, but – more interestingly – that they also determine the representational space of an item's memory trace. Specifically, items that are encoded at higher task-relevant distances are remembered better, possibly reflecting effects of contextual deviance[38–41] and/or distinctiveness[42–44] (see Discussion).

Employment of graded and metric representational spaces following binary ratings

In Studies I and II, the representational spaces of item-specific memory traces were defined by explicit distance ratings along graded similarity scales. Next, we tested whether representational distances during encoding also affected memories if participants performed binary decisions in a forced-choice paradigm (Fig. 3a). During encoding in Study III, participants first saw a target image and then indicated which of two probe images was more similar to the target image. This decision was again based on either perceptual or conceptual

features. We observed high levels of accuracy in selecting the correct choice, with an even higher accuracy for conceptual vs. perceptual ratings (conceptual: M = 0.98, SD = 0.02; perceptual: M = 0.96, SD = 0.03; paired t-test across participants: $t_{27} = 4.23$, $p = 0.0002$; the effect size, as measured by Cohen's d, was d = 0.82).

In order to identify the representational distances that guided participants' judgements in the absence of graded similarity ratings, we employed DNNs to extract the similarities between pairs of objects in different representational formats (Fig. 3c-d). We used a cDNN (AlexNet[45]) and a dNLP (Google's universal sentence encoder[46]; USE) to assess representational similarities in perceptual, higher-order perceptual, and semantic representational formats (Fig. 3c; see Supplementary Fig. 3 for more details). We first confirmed the correspondence of representational distances in these DNN formats to explicit similarity judgements, using the data from Studies I and II. We defined recruitment of low-level perceptual, higher-order perceptual and semantic representational spaces as averaged similarities with representations in convolutional layers, fully connected layers, and the USE, respectively (see Methods for more detail and Supplementary Fig. 3c for layer-wise correlations). Using these three formats in a linear mixed model, we found that distances in low-level perceptual space ($z = 4.54$, $\chi^2_{(1)} = 20.67$, $p < 0.001$; likelihood ratio test) and higher-order perceptual space ($z = 11.40$, $\chi^2_{(1)} = 129.51$, $p < 0.001$; likelihood ratio test) but not semantic space ($z = 1.87$, $\chi^2_{(1)} = 3.53$, $p = 0.06$; likelihood ratio test) predicted perceptual ratings (Supplementary Fig. 3b). Reversely, distances in higher-order perceptual space ($z = 15.80$, $\chi^2_{(1)} = 247.89$, $p < 0.001$; likelihood ratio test) and semantic space ($z = 7.30$, $\chi^2_{(1)} = 53.30$, $p < 0.001$) but not low-level perceptual space ($z = 1.76$, $\chi^2_{(1)} = 3.10$, $p = 0.07$; likelihood ratio test) predicted conceptual ratings (Supplementary Fig. 3b).

These results suggest that the task-relevant binary ratings in Study III can be explained by DNN-based estimates of distances in distinct representational spaces. For each image pair, we thus computed the DNN similarities between the target stimulus and the correct choice image and between the target stimulus and the incorrect lure image (Fig. 3d). The difference between these similarities was considered a proxy for distances in task-relevant representational spaces underlying the decision (see Supplementary Fig. 4 for more details). Similar to Studies I and II, these distances predicted response times: Response times during perceptual choices were predicted in a linear mixed model with a random participant intercept using distances in low-level perceptual space ($z = -3.90$, $\chi^2_{(1)}=15.20$, $p = 0.001$; likelihood ratio test) but not higher-order perceptual space ($z = -1.29$, $\chi^2_{(1)} = 1.67$, $p = 0.196$) or semantic space ($z = 0.46$, $\chi^2_{(1)}= 0.22$, $p = 0.639$; likelihood ratio test). By contrast, response times during conceptual choices were predicted by distances in both higher-order perceptual space ($z = -4.91$, $\chi^2_{(1)} = 24.00$, $p < 0.001$; likelihood ratio test) and semantic space ($z = -2.68$, $\chi^2_{(1)} = 7.21$,

$p = 0.007$; likelihood ratio test), but not low-level perceptual space ($z = 0.49$, $\chi^2_{(1)} = 0.24$, $p = 0.621$; likelihood ratio test) (Fig. 3e).

Next, we again applied linear mixed models to test whether memory traces were embedded into task-relevant representational spaces, i.e., whether task-relevant distances during encoding predicted memory strength. In Study III, we refined the recognition task to separately address different recall demands either based on generalized, or pattern-completed, information ("Did you see an image from this category during encoding?") or based on highly differentiated, or pattern-separated, information ("Did you see exactly this image during encoding?") (Fig. 3b). Consistent with Studies I+II, we did not find an effect of encoding space on memory strength when using the average confidence of both the specific and general memory test for old images as predictors in a linear mixed model ($z = -0.43$, $\chi^2_{(1)} = 0.19$, $p = 0.6650$; likelihood ratio test; Fig. 3f). Importantly however, for perceptually encoded items, representational distances in higher-order perceptual space ($z = 3.91$, $\chi^2_{(1)} = 5.24$, $p = 0.022$; likelihood ratio test), but not low-level perceptual space ($z = 2.91$, $\chi^2_{(1)} = 0.55$, $p = 0.460$; likelihood ratio test) or semantic space ($z = 3.71$, $\chi^2_{(1)} = 2.06$, $p = 0.150$; likelihood ratio test; Fig. 3g) predicted memory strength (general and specific memory tests combined; see Supplementary Fig. 5 for separate effects in the general and the specific memory test). For conceptually encoded items, representational distances did not predict subsequent memory strength (all $p > 0.597$).

Together, these results indicate that task-relevant representational distances can affect recognition memory performance, suggesting that memory traces of individual items are embedded in different representational spaces depending on encoding instructions. The linear relationship between task-relevant representational distances during encoding and memory strength (i.e., the accessibility of item-specific information for recognition memory judgements) suggests that memory traces involve coding schemes with metric cognitive maps in specific task-relevant formats (see Discussion).

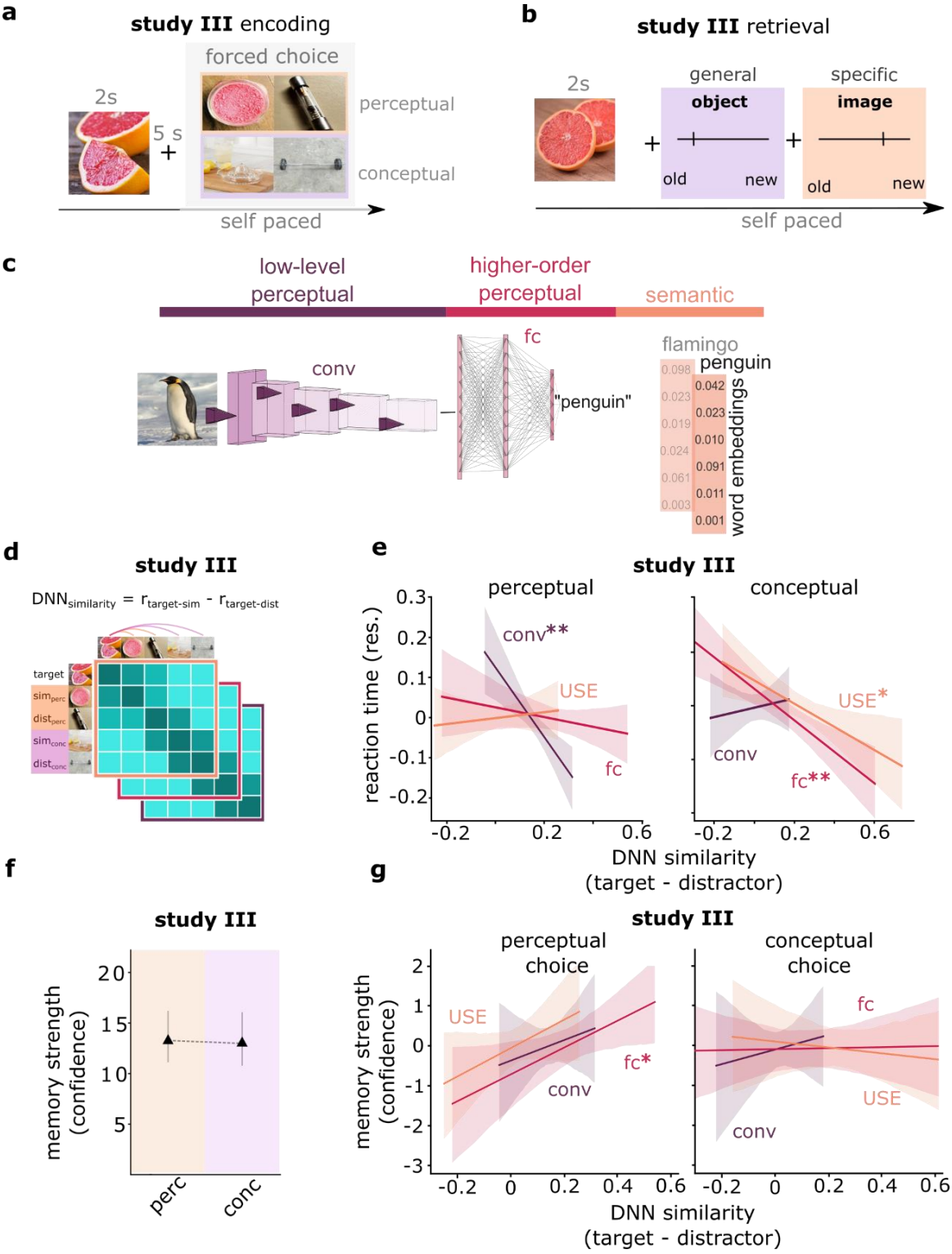


Figure 3. Memory traces with DNN-based representational distances following binary similarity judgments. a) In Study III, participants performed forced-choice similarity ratings during encoding. Given a target image (blood orange), they chose a conceptually (lemon press) or perceptually (pink power pod) similar image from two options. b) Participants first performed a general retrieval task (indicating whether a given image matched a previously presented category) and then a specific retrieval task (indicating whether the image matched the previously presented exemplar). c) Graded representational distances in low-level perceptual formats, higher-order perceptual formats and a semantic format were derived from a convolutional DNN and the universal sentence encoder (USE), respectively. d) We computed the difference in DNN similarities between target image and similar (correct) choice vs target and distractor (incorrect) choice, separately for each DNN level. e) Faster

reaction times are predicted by higher DNN similarity differences in task-relevant representational formats: reaction times in the perceptual task correspond to DNN similarities in convolutional layers (conv), while reaction times in the conceptual task correspond to DNN similarities in fully-connected (fc) layers and in the semantic network (USE). f) Encoding task does not influence memory strength. g) Task-relevant DNN-derived representational distances predict memory strength. Data are visualized after removing participant-wise estimated random effects. 95% confidence intervals, error bars indicate SEM. *, p<0.05; **, p<0.01.

Effects of representational formats on different task demands during memory retrieval

We next explored the functional relevance of the different representational spaces for subsequent retrieval demands. Based on previous studies[47,48], we hypothesized that encoding in a conceptual space may be particularly beneficial for retrieval of generalized, perceptually invariant categorical information[49–51]; reversely, encoding in perceptual spaces may be relevant when retrieval requires the rejection of similar but not identical exemplars (i.e., pattern separation).

We first analyzed whether the different retrieval instructions in Study III (general/specific) shifted memory decisions (Fig. 4a). Indeed, in the general compared to the specific retrieval task, hit rates were significantly higher (t_{27} = 11.55, p < 0.001; paired t-test; the effect size, as measured by Cohen's d, was d = 1.50) and correct rejection rates of new concepts were significantly lower (t_{27} = -12.35, p < 0.001; paired t-test; the effect size, as measured by Cohen's d, was d = 2.59), suggesting a more liberal recognition memory threshold. Note that responses to new exemplars could not be directly compared because they required "old" ratings in the general task but "new" ratings in the specific task.

Next, we investigated whether performance in the general and the specific memory task depended on task-relevant encoding spaces (Figure 4b). We used a linear mixed model to predict memory performance using retrieval task (general/specific) and encoding space (perceptual/conceptual) as predictors with a random participant intercept and including an interaction between retrieval task and encoding space. To test for significant differences, we then used a likelihood ratio test between the full model containing the main effect and the interaction term and the reduced model excluding the interaction. Indeed, we found an interaction between retrieval task-demands and encoding space (z = 2.69, $\chi^2_{(2)}$ = 7.25, p = 0.007; likelihood ratio test). Conceptual encoding improved memory in the general retrieval task (z = 3.04, $\chi^2_{(1)}$ = 9.27, p = 0.0023; likelihood ratio test) but not in the specific retrieval task (z = 1.28, $\chi^2_{(1)}$ = 1.66, p = 0.197; likelihood ratio test; Fig. 4b, left panel). In addition, we found a significant interaction for new exemplars (z = 2.53, $\chi^2_{(2)}$ = 6.44, p = 0.011; likelihood ratio test): Conceptual encoding benefitted generalized memory decisions, i.e., hit rates towards new exemplars were higher than for perceptually encoded items (z = 4.53, $\chi^2_{(1)}$ = 20.52, p < 0.0001; likelihood ratio test), while performance in the specific retrieval task did not depend on

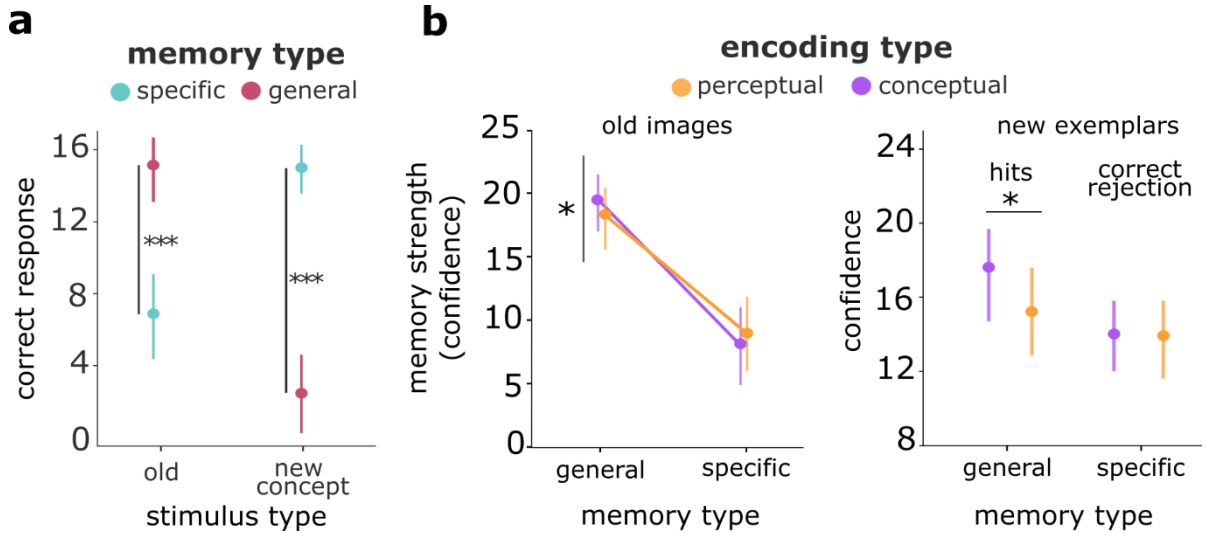


Figure 4. Conceptual encoding improves generalized retrieval and does not impair specific retrieval. a) Different response criteria depending on retrieval tasks: General retrieval leads to higher hit rates but lower correct rejection rates of new concepts. b) Left: conceptual encoding improves memory strength during general but not specific retrieval. Right: conceptual encoding facilitates correct recognition that new exemplars belong to previously presented categories during general retrieval and does not influence performance during specific retrieval. Data are visualized after removing participant-wise estimated random effects. Error bars indicate SEM. *, p<0.05, ***, p<0.001

encoding instructions ($z = 0.47$, $\chi^2_{(1)} = 0.22$, $p = 0.6354$; likelihood ratio test; Fig. 4b, right panel). Additional analyses indicated that employment of generalized information (in the general retrieval task) specifically benefitted from lower representational distances in higher-order perceptual formats, while rejection of new concepts was impaired by lower distances in this format in both retrieval tasks (Supplementary Fig. 6).

These results indicate that generalized information at test was more readily available for conceptually encoded images for which participants presumably relied more on gist-like representations. In other words, conceptual encoding created memory traces with generalizable representations that facilitated the correct recognition that new exemplars belong to previously presented categories. Thus, encoding stimuli into conceptual spaces improves both the accessibility of their memory traces (higher memory strength of previously presented exemplars) and their employment for generalized memory decisions (better recognition that new exemplars are members of a previously presented category). Interestingly, this does not come at the expense of less perceptually distinct representations since performance in the specific memory test was not reduced, suggesting that conceptually encoded items are embedded into a “deep” memory trace with multiple different formats (even though task-irrelevant spaces do not seem to involve metric coding schemes; see Discussion).

Selective strengthening of task-relevant formats during consolidation

Our results presented thus far indicate that memory traces contained encoding spaces in distinct task-relevant representational formats, but also reveal an asymmetry, because

conceptual encoding spaces benefitted both general and specific retrieval, whereas perceptual encoding spaces did not offer any particular advantage. Since memory consolidation has been suggested to strengthen conceptual representations[28], we explored its influence on the representational formats of memory traces. We employed TMR and tested whether cueing with a sound that was presented during either perceptual or conceptual encoding selectively enhanced the corresponding representational format (Figure 5a).

We first analyzed how TMR affected the memory strength of previously presented items, using a multi-level model with encoding task and cueing as predictors and participant as random effect (Fig. 5b). We observed a main effect of cueing ($z = -2.12$, $\chi^2_{(1)} = 4.53$, $p = 0.0333$; likelihood ratio test), indicating generally better memory for cued vs. uncued items, but also a significant interaction ($z = 4.38$, $\chi^2_{(2)} = 19.23$, $p < 0.001$; likelihood ratio test), indicating different cueing effects on perceptually vs. conceptually encoded items. Post-hoc tests showed that cueing only improved the memory strength of conceptually encoded items ($z = -4.38$, $\chi^2_{(1)} = 19.23$, $p < 0.001$; likelihood ratio test) but not of perceptually encoded items, for which even a trend in the opposite direction was observed ($z = 1.88$, $\chi^2_{(1)} = 3.54$, $p = 0.06$; likelihood ratio test).

Since our results thus far show that memory traces are selectively embedded in task-relevant representational spaces, but also suggest that information in irrelevant formats remains present after encoding, further strengthening one of these formats may come at the expense of other formats: Even though conceptual encoding by itself does not impair specific retrieval of perceptual information (Fig. 4b), one may hypothesize that further strengthening of conceptual formats during consolidation reduces the perceptual specificity of item representations, increasing false alarms to novel exemplars (i.e., lures) of previously encoded categories. We thus implemented a linear mixed model for memory ratings of similar lures with encoding condition and cueing as predictors (Fig. 5c). This model indeed revealed a significant interaction ($z = -3.32$, $\chi^2_{(2)} = = 13.11$, $p = 0.0003$; likelihood ratio test), showing different effects of cueing on correct rejection of similar lures for conceptually vs. perceptually encoded items. Follow-up tests showed that cueing of conceptually encoded items deteriorated the identification of similar lures ($z = 2.38$, $\chi^2_{(1)} = 5.70$, $p = 0.017$; likelihood ratio test), while cueing of perceptually encoded items improved similar lure identification ($z = -2.77$, $\chi^2_{(1)} = 7.66$, $p = 0.0056$; likelihood ratio test). Consistent with our results in Study III, conceptual encoding by itself (i.e., without cueing) did not deteriorate lure detection ($z = -0.35$, $\chi^2_{(1)} = 0.12$, $p = 0.726$; likelihood ratio test). These results indicate that cueing can selectively increase either conceptual representational formats (improving recognition of old items but deteriorating rejection of lures) or perceptual formats (with no effect on later memory but more accurate rejection of lures).

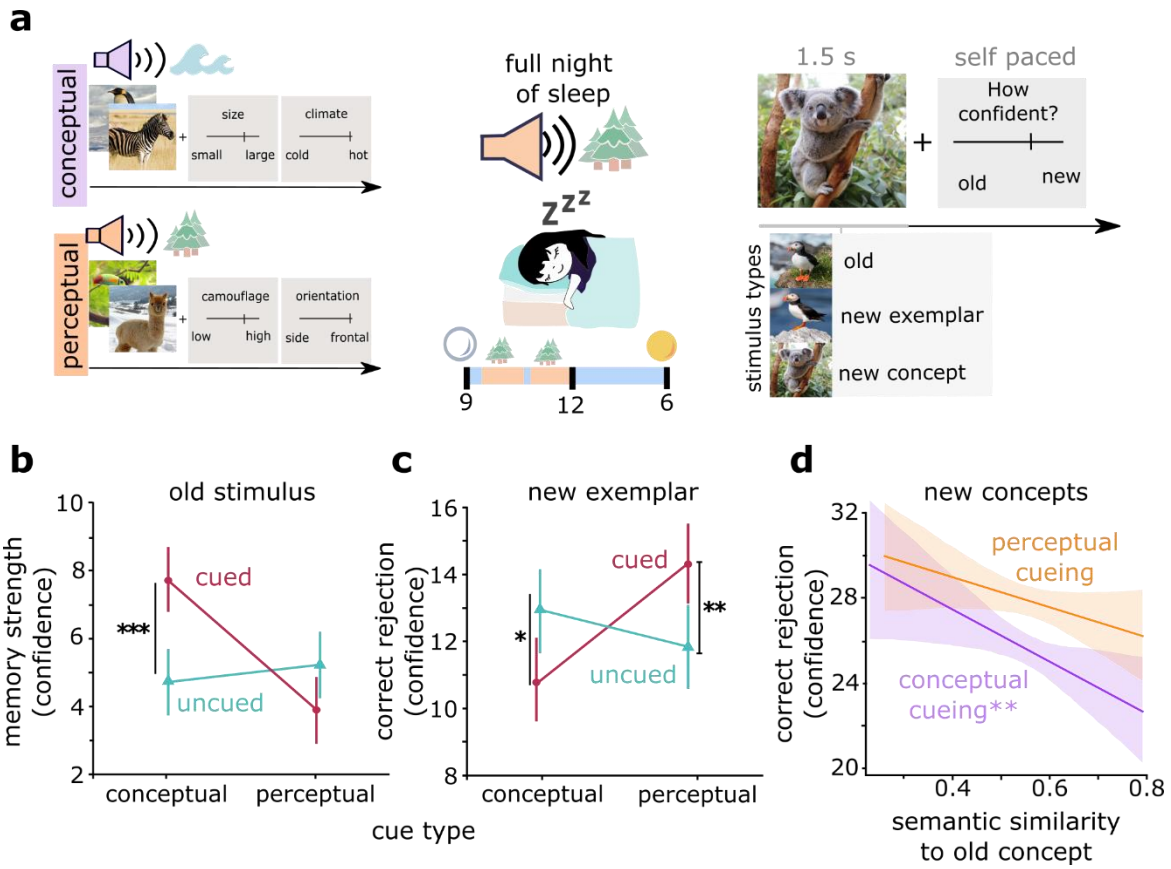


Figure 5. Strengthening of task-relevant and weakening of task-irrelevant representational formats after targeted memory reactivation (TMR). a) TMR design: presentation of condition-specific auditory cues during conceptual and perceptual encoding blocks (e.g., coast sound and forest sound). One of these sounds was played again during the first half of the subsequent night, followed by a graded recognition memory test as in Study I. b) Cueing of conceptual but not of perceptual formats increases memory strength. c) Cueing of perceptual formats improves and cueing of conceptual formats impairs correct rejection of new exemplars. d) Higher semantic similarities of new concepts to previously presented concepts impairs correct rejection following conceptual but not perceptual cueing. Data are visualized after removing participant-wise estimated random effects. 95% confidence intervals, error bars indicate SEM. *, p<0.05, **, p<0.01, ***, p<0.001

Does cueing of conceptually encoded items also impair the ability to identify novel concepts if they are sufficiently similar to previously encoded categories? We used the semantic similarities, quantified by the dNLP, of novel concepts to all previously encoded concepts as a predictor in a linear mixed model (Fig. 5d) and indeed found that cueing conceptually encoded items impaired identification of semantically related novel concepts ($z = -2.65$, $\chi^2_{(1)} = 7.01$, $p = 0.008$; likelihood ratio test). This effect was not found after perceptual cueing ($z = -1.60$, $\chi^2_{(1)} = 2.56$, $p = 0.11$; likelihood ratio test). As a control, we confirmed that the influence of low-level perceptual and higher-order perceptual similarities on the identification of novel concepts did not differ between cueing conditions (Supplementary Fig. 7).

These results indicate that representational formats of memory traces are not fully determined by the initial task-relevance during encoding, but that they can be further and selectively strengthened during subsequent consolidation stages. Our findings further show that depending on the task demands to either recognize previously presented items or to reject similar lures, these consolidation processes may exert either beneficial or detrimental effects.

Discussion

Generative memory theories[4,52,53] propose that memories are inherently constructive: They suggest that memory traces are rapidly transformed from perceptual into conceptual/semantic representational formats, and that missing information is supplemented from prior knowledge or schemas at the time of retrieval[3,54]. Other frameworks propose that memory traces contain richer representations in multiple formats whose relative accessibility can be flexibly determined by various factors during encoding, consolidation, and retrieval[5,55]. The aim of our study was to systematically investigate the role of representational formats in memory traces and to investigate whether transformations of a memory trace are indeed unidirectional, such that perceptual formats are irreversibly lost in favor of semantic information, or whether multiple formats coexist. Our study yielded three main results: First, we showed that memory traces consist of distinct representational formats that are determined by task demands during encoding and affect the accessibility of information during subsequent retrieval. The linear relationship between representational distinctiveness and memory strength suggests that stimuli are embedded into metric encoding spaces that are determined by objective item features which correspond to different formats of DNNs. Second, we found asymmetric functional roles of these different encoding spaces for subsequent retrieval demands: While conceptual encoding benefits retrieval during generalized task conditions, it does not impair performance during specific retrieval, suggesting that perceptual information is still maintained in a “deeper” memory trace. Finally, we observed that consolidation selectively strengthens task-dependent representational formats and weakens task-irrelevant formats, leading to “shallower” memory traces.

While we did not find a pronounced effect of encoding format on subsequent memory, we observed robust evidence for a role of task-relevant representational spaces. Specifically, the distances of task-relevant but not irrelevant dimensions predicted the memory strength of both perceptually and conceptually encoded items. Thus, task-demands[7] but not the mere depth of processing shapes which features are prone to decay, and which are preserved and even strengthened to support later memory. Interestingly, we found better memory if stimuli were encountered at larger representational distances in their relevant encoding space. How can we explain the higher accessibility of stimuli encoded at higher task-relevant distances? First,

one may assume that strengthened memory for more dissimilar items could be caused by higher prediction errors[56–58]: While the current reference position e.g. of a shark in the "approachability" space might lead to the expectation of further dangerous animals, the presentation of a chinchilla would lead to greater prediction errors. However, this explanation cannot account for our finding that not only the second, but also the first image in pairs with larger distances was remembered better. For the same reasons, contextual deviance (as in the von Restorff effect[41]) is less likely the main factor driving this result[59,60]. Alternatively, higher dissimilarities may promote processes of pattern separation[61] which may facilitate the disambiguation of similar images during retrieval. However, we did not find that higher task-relevant encoding distances facilitated the correct rejection of similar lures. Previous studies showed that the hippocampus encodes representational distances in task-relevant cognitive maps[62,63]. Here we extend these findings by showing that these distances are actually relevant for subsequent memory, and thus contribute to theories bridging spatial navigation and episodic memory[20,64]. Specifically, we found that representational distances were linearly related to the accessibility of items for later memory, even if items were not originally encoded using graded distance ratings (Study III). This suggests that memory traces involved cognitive maps with interval-scaled (metric) rather than rank-based (ordinal) coding schemes[65], in line with the original concept of cognitive maps to promote shortcuts during physical navigation[66].

Notably, while all three studies presented here provide converging evidence for a selective strengthening of task-relevant but not task-irrelevant representational formats, we also found consistently that conceptual encoding by itself did not come at the expense of reduced identification of similar lures. This suggests that the benefit of conceptual encoding may not (only) consist in deeper encoding and faster integration with pre-existing semantic knowledge, but in the formation of "deeper" and more versatile memory traces that can be flexibly employed for both specific and generalized retrieval demands. Even though the accessibility of perceptual information in conceptually encoded memory traces does not depend on (in this case, task-irrelevant) perceptual distances and is thus not organized in a "metric" cognitive map (see above), it still remains available even following a night of sleep, as in Study II. Future studies will need to investigate whether more extended consolidation processes across longer time periods promote the formation of "shallower" traces in purely conceptual/semantic formats, leading to a more pronounced decline in the availability of perceptual information.

Interestingly, and to our knowledge for the first time, we observed detrimental effects of conceptual as compared to perceptual encoding on lure identification following TMR, suggesting that enhanced consolidation processes induce a "pruning" of memory traces that removes information in task-relevant representational formats. Previous studies demonstrated that cueing during sleep is associated with reactivation of the associated memories[67–71]. After

cueing conceptually encoded items, participants relied more on gist-like representations, causing a decreased performance in the rejection of both visually similar lures and semantically similar concepts. These findings are in line with results that reactivation during sleep facilitates generalization processes of features that are shared across several encoded contents[72–74], which in turn increases false alarm rates for similar lures[75]. However, while this may be interpreted as evidence for a unidirectional transformation of memory traces from perceptual into conceptual formats during consolidation, we also found that cueing of perceptually encoded items improved the correct rejection of visually similar lures. This suggests that TMR – and possibly, memory consolidation and sleep-related reactivation more generally – selectively enhances task-relevant[31,76] and reduces task-irrelevant representational formats, rather than generally transforming perceptual into conceptual formats. Thus, while multiple representational formats coexist in memory traces after encoding, these formats can be selectively strengthened and weakened by consolidation processes[5].

Our results are solely based on behavioral measures, emphasizing the utility of behavioral paradigms during (arguably) ecologically valid tasks for inferring complex properties of human cognitive representations[77]. Nevertheless, we propose that follow-up studies should aim to investigate these representations more directly using neuroimaging and/or electrophysiological recordings. Specifically, previous studies emphasized a role of the hippocampus for representing multidimensional abstract spaces[20,63,78], and it would be interesting to see whether hippocampal circuits can be flexibly adapted to represent the respective task-relevant formats due to their high degrees of mixed selectivity[79], or whether different formats are represented in different hippocampal regions e.g. along its long-axis[80–82]. Moreover, while previous studies investigated representations in newly formed abstract spaces[83–85], participants in our study performed similarity and memory judgements in pre-existing spaces of semantic and perceptual relationships which may involve interactions of hippocampal distance representations with distinct regions along the VVP. Finally, neuroimaging studies are ideally suited to investigate “dormant” memory traces during states without direct behavioral readout such as awake resting periods, sleep, or while concurrent cognitive tasks are performed, and could thus explore the mechanisms underlying the “pruning” of memory traces during consolidation.

Our findings demonstrate the heuristic potential of DNNs to investigate the representational spaces underlying similarity judgements and the selective embedding of these spaces into memory traces. Previous studies already showed that DNNs can be used as a proxy for perceptual and semantic formats during memory tasks[18,23,24,86,87] and behavioral similarity judgements[87,88], and our results indicate that the same models can account for similarity judgements during encoding and their impact on subsequent memory. Importantly, we found

that distances from cDNNs and from a natural language model contributed differently: Whereas distances based on perceptual cDNN features predicted perceptual ratings, reaction times, and memory strengths, distances of deeper cDNN layers and of a semantic model predicted behavioral outcomes related to conceptual formats. These results emphasize the usefulness of DNN models as proxies of different representational formats – even in behavioral outcomes – and contribute to the idea of multiple formats in a memory trace.

Taken together, we propose that both perceptual and conceptual representational formats coexist in a multi-layered memory trace, with each layer represented by a different format. Various factors such as attention, prior knowledge and task-demands may determine the accessibility of a specific format and can thus account for the apparent transformation of memories over time and contexts.

Materials and Methods

Sample

We tested adult healthy students from the Ruhr University Bochum (Germany). Participants signed their informed consent and received study credits as compensation. The study was approved by the ethics committee of the Faculty of Psychology, Ruhr University Bochum, and was conducted according to the declaration of Helsinki. In study I we tested 47 participants (age M = 22.9, SD = 4.0, 38/9 female/male). In study II the sample consisted of 55 participants (age M = 23.05, SD = 3.57, 44/11 female/male) after excluding one participant due waking up during the night and recognizing the sound played. In study III we tested 28 participants (age M = 23.17, SD = 3.22, 16/12 female/male), after excluding 3 participants due to technical issues. Analysis of DNN contributions to prediction of encoding distances was done on the combined data from study I and II (n=102).

Behavioral Tasks

Data for all three studies were collected online. We implemented the tasks for study I and II using jsPsych[89] and Cognition (https://www.cognition.run/). In study III we implemented the task using Psychopy3[90] and Pavlovia (https://pavlovia.org/).

We chose images from the THINGS database[91], a large database of natural images from different categories (e.g. animals, objects, food), specifically built to make DNN training sets more naturalistic and applicable to neuroscience. represent an ecologically valid reflection of human-relevant everyday objects. For study I and II, we used animal images only to avoid effects of animate/inanimate differences during later similarity analyses. We excluded images containing multiple animals or containing humans. Since the encoding phase of study III consisted of predefined forced choice trials, the conceptual and perceptual triplets (target and

two choice options) showed not only animals, but equal amounts of objects from other categories contained in the THINGS database, namely food, tool, container, vehicle, and a compilation of individual objects from miscellaneous categories (e.g. musical instruments, electronic devices, toys). In each participant, the amount of target objects from each category was kept equal within encoding blocks and across perceptual and conceptual conditions. During recognition, new concepts were selected from the same pool of categories and in equal amounts.

Study I and II consisted of two parts, an incidental encoding task (Fig. 1b) and a recognition task the day after (Fig. 1c). During encoding participants rated conceptual or perceptual similarities of a series of image pairs drawn from different categories. In each trial we consecutively presented two images for 1.5 seconds each followed by two self-paced ratings, which required either conceptual or perceptual similarity judgements. Participants rated similarities of the last presented image in relation to the first image on a visual rating scale. Conceptual judgements included "size" (Is the animal smaller or larger?), "climate" (Does the animal live in a colder or a hotter climate?) and "approach" (Is the animal more approachable or less?). Perceptual judgements included "orientation" (Is the animal more facing towards the viewer or more away from the viewer?), "colors" (Does the image contain warmer or colder colors?) and "camouflage" (Is the animal concealed more or less?). Since we wanted to examine if perceptual or conceptual formats, or the combination of both, would strengthen subsequent memory, we divided the six rating dimensions into three different block types. Each type consisted of two dimensions from either the same or different conditions (perceptual, conceptual, mixed). We presented 336 images of 84 animals in total, with each animal being shown in one of the three block types only (e.g., the 4 different images of giraffes were all shown in perceptual blocks). Block types of each animal and image pairing were randomly shuffled between participants.

In Study II we investigated the role of consolidation on the transformation of representational formats. We used the same design as in study I but paired conceptual and perceptual ratings with condition-specific sounds (Fig. 5a). During each encoding block we continuously played one of two distinct nature context audios[92]. One audio was composed of ocean waves and sounds of the sea. The other audio consisted of forest sounds like the brushing of leaves in the wind, frogs croaking and crickets chirping, excluding bird sounds to avoid interference with our animal classes. One of the two sounds was played again during post-encoding sleep in order to reactivate memories associated with the sound[93,94]. Each audio file was randomly assigned to be played either during conceptual or perceptual encoding blocks. We decided to exclude mixed rating blocks, as there would be no clear conclusion as to how consolidation of both rating dimensions should impact subsequent memory performance while at the same time

this would have increased the number of trials to cue during sleep. In the night following the encoding session, participants were asked to play a custom-made audio file through their smartphone, starting the file when they would start trying to fall asleep for the night. In standardized studies applying targeted memory reactivation, EEG data would be monitored on-line to specifically play sounds during slow wave sleep[31,95]. Due to practical reasons and to ensure contact-free participation in our study during the Covid 19 pandemic, we decided to play an audio file timed in accordance to the expected higher amount of slow wave sleep in the first half of the night[96–98]. Audios were composed to play either the perceptual or conceptual sound during the first half of sleep with long periods of slow wave sleep, highly relevant for consolidation[95,97]. The audio file started with a masking white noise to habituate participants to the volume to avoid waking them up due to the sudden onset of the actual cue playing during sleep. Sixty minutes after audio onset, the cue was played for 30 minutes, followed by 60 minutes of masking noise. The cue was then played again for 30 minutes. If participants had been awake for longer periods or had to use the bathroom until 2 a.m., they were instructed to restart the audio, though no participant reported waking up prior to this time. The next morning, participants were asked whether they heard any specific sounds during sleep, woke up or stayed awake for longer time periods. If participants recognized the sounds being played during sleep or if they did not have a full night of sleep between encoding and recognition, we excluded these participants from later analysis (n=1). We also excluded trials for study I and II with reaction times greater than 150 seconds.

The day following encoding, participants performed a recognition memory task which was identical for studies I + II (Fig. 1c). We presented the same images again (“old”), mixed with a total of 336 new images. Novel images were either unknown images of previously presented animal species ("new exemplars") or images of animal species that had not been presented before ("new concepts"). Each image was presented for 1.5 seconds. Afterwards, participants indicated their confidence of having seen the image before on a continuous rating scale from old to new. All 336 old images, 168 new exemplars and 168 new concepts were shown in random order.

To examine the role of binary rather than gradual similarity ratings and of different retrieval demands, we implemented different encoding and retrieval tasks in study III (Fig. 3a-b). Instead of subjectively rating the similarity of image pairs during encoding, participants performed forced-choice similarity judgements of triplets (three serially presented images). Each triplet began with a 2s presentation of the target image, followed by a 5s maintenance period during which a fixation cross remained on screen. After maintenance, two horizontally arranged images were presented on either side of the fixation cross (left and right choice option). Participants were instructed to select the image that was more similar to the initially presented

target on either perceptual or conceptual dimensions. Similar to Study I and II, participants performed blocks of the encoding task in which either perceptual (image) or conceptual (object) features were relevant for performing the task.

In order to ensure reliable and valid responses across participants, we carefully constructed trials for each target image where either a perceptually or conceptually similar option was shown together with an unrelated distractor. Stimuli were selected based on similarity predictions from the cDNN. For each of the 120 target images, we manually selected images to form a conceptual and perceptual triplet by identifying images from the THINGS database which predominantly shared feature similarities with the target on either early (convolutional) DNN layers (perceptual, e.g.: similar colors, shapes; example: blood orange - pink powder) or later (fully-connected) DNN layers (conceptual, e.g.: similar object category; example: blood orange - lemon press). Accordingly, for each triplet we selected distractor images that did not share the respective features with the target image. To ensure a meaningful difference in similarity of target and choice options, we quantified the difficulty of each triplet by computing the difference in the similarity between the target and the matching stimulus (r_{sim}) and the similarity between the target and the distractor (r_{dist}). For the distribution of scores on each visual and semantic DNN layer see Supplementary Figure 4. We restricted our analysis to triplets with > 75% accuracy across participants, which led to exclusion of n=5 perceptual triplets.

After successfully performing the forced choice task, participants received monetary reward feedback of 2 cents (low), 10 cents (high). Reward conditions did not elicit behavioral differences across conditions in neither encoding or recognition nor their congruency in encoding-recognition and are thus not further discussed. For all analyses of study III encoding, N = 3,161 trials were considered after excluding trials with incorrect forced choice response (n = 86) and trials with reaction times exceeding four standard deviations above the participants' average (n=44).

Similar to studies I+II, participants engaged in a recognition memory test on the following day (Fig 3b). During this test, 120 old images, 120 new exemplars and 120 new concepts were presented for 2 seconds each. In each trial, participants consecutively indicated their confidence of having seen the object (general retrieval) and having seen the specific image (specific retrieval) using two separate scales. Participants indicated general memory before specific memory, similar to the procedure in Davis et al. (2021). Note that while for old images (old concept, old image) and new concepts (new concept, new image), correct responses in the two retrieval tests were on identical sides of the slider scale, this differed for new exemplars, which required “old” responses in the general retrieval task but “new” responses in the specific retrieval task.

As for the encoding task, we excluded trials with reaction times exceeding four standard deviations above the participants' average reaction time (n=120 perceptual trials, n=120 conceptual trials), resulting in a total trial number of N= 9,494 recognition trials across all participants.

Linear mixed models

We analyzed all data from the three studies using linear mixed models (MLM) of the statsmodels package[99] in Python 3.8. To determine the influence of different factors on behavioral performance, we set up a full model containing all fixed effects and participant as a random effect. We then use a likelihood test to determine whether the exclusion of the factor of interest significantly reduces the explained variance compared with the full model. For main effects with one predictor the reduced model included a constant as fixed effect (1). For main effects with two or more predictors the reduced model contained all other main effects except the effect of interest. For interaction effects, the full model contained all main and interaction effects. However, main effects from interaction models were not interpreted. For each fixed effect, we report z-statistics from the initial estimation of the model and the respective likelihood derived from model comparison ($\chi 2$) and its significance level (p). Beta weights resulting from the full models are shown in the Supplementary Table 1 (for the main results) and Supplementary Table 2 (for analyses from the Supplement). For post-hoc estimation of condition differences we ran separate models on data split by the factor of interest, and then compared the likelihood of the reduced model with only a random intercept. For visualization purposes of individual fixed effects, we plotted the residuals of mixed models missing the fixed effect of interest. Finally, individual models were used to determine the behavioral performance in conceptual vs perceptual encoding tasks as well as memory performance of old images, new exemplars, and new concepts during recognition.

Relevant versus irrelevant distances

To compare the effect of relevant and irrelevant distances in Study I+II, we first constructed individual one-dimensional spaces for each dimension by computing a least squares solution on a sparse matrix of all image ratings on the respective dimension over all participants (Figure 2a). For this, we first created a design matrix M with dimensions of the number of ratings by the number of images containing zeros. Next, M was filled for each rating with 1 or -1 indicating if an image was first (1) or second (-1) in the rating trail. The design matrix M was then converted into a compressed sparse column matrix (CSC). A least-squares solution was then used to solve the linear system of equations on our design matrix M given the ratings done by participants. Two least squares solutions can then be used to compute the distance of each image pair in the respective 2D space. We then use these spaces to extract the distances of

image pairs on dimensions that the participant did not rate the image pair on, thus creating an "irrelevant" distance as an average over all spaces. The irrelevant distance in this case thus corresponds to a distance that was not relevant for the participant during the task. In contrast to this, we define the relevant distance as the similarity judgment of each image pair that the participant performed during the task. For example, when a participant rated the similarity of a panda and a giraffe on the dimensions of climate and size (rated distance), we then use the average distance from the least square solutions of all other spaces (climate + approachability; size + approachability; color + camouflage; orientation + camouflage; color + orientation) as the irrelevant distance.

DNN similarities

We used a pre-trained convolutional deep neural network, "AlexNet"[45], as implemented in the Caffe framework[100] in Python 2.7, and abstract semantic representations from the Google Universal Sentence Encoder[46] (USE) using tensorflow[101] in Python 3.8. (Fig 3c). To analyze cDNN similarities, we first generated features for each layer of the AlexNet for all images. In each layer we averaged across the spatial dimension keeping only one value per feature. We then correlated the features of each image with all other images using spearman's *rho* correlation resulting in 8 image by image correlation matrices, one for each layer. These correlation maps represent the similarity of each image to all other images across the layers of the cDNN starting with similarities based on early visual information in convolutional layers and concluding with higher-order visual information of fully connected layers (Fig 3c). We averaged the correlation maps of convolutional and fully connected layers to simplify later interpretation of differences concerning behavioral performance and reducing the number of statistical tests to correct for. Thus, we obtained two correlation matrices from AlexNet, which served as our markers of low-level perceptual representations (convolutional layers) and higher-order perceptual representations (fully connected layers), while the USE model contains abstract semantic feature formats. We computed word embedding vectors for all concept labels retrieved from the THINGS database and correlated each concept embedding vector with all others to generate a concept-by-concept correlation matrix.

Pairwise image similarities from perceptual and semantic DNNs were used to quantify the degree of overlap between pairwise images given specific features (Fig 3d). This was done for all images of a triplet in study III, all pairs of old images and their corresponding new exemplars, as well as to estimate the distance of new concepts to the bulk of all previously seen concepts by fetching the highest similarity to all old concepts (i.e., smallest distance).

Similarity matrices were also used to study whether DNN similarities reflected rated distances in study I and II (Supplementary Fig. 3). We expected the effect to be dependent on the rated

dimension (conceptual vs. perceptual) with higher contributions of early (convolution) cDNN layers to perceptual ratings, and of late (fully connected) cDNN layers and the USE to conceptual ratings. This analysis was done on the layer-wise correlation maps to investigate the effect in more detail across the DNN gradient. For each layer we performed a one-sided paired t-test (scipy package[102]) on fisher-z transformed average correlations between participant ratings of perceptual and of conceptual dimensions (conceptual, perceptual, mixed for Study I; conceptual and perceptual for Study II) and the similarities of the respective cDNN layer and the USE, resulting in 9 paired t-tests. We then corrected for multiple comparisons using FDR-correction. We computed the difference between perceptual and conceptual similarities in each layer to plot layer-wise relations to similarity judgements. Next, we used layer-averaged, and fisher-Z transformed similarity matrices as predictors in mixed models for rated distance during encoding and for memory strength during recognition.

For study I and II, we computed the similarity of a new exemplar or new concept to all images previously shown during encoding on conv and fc layers for perceptual similarity and the similarity of a new concept to all previous concepts and extracted the highest similarity for later analyses. For study III, we computed the similarity of a new exemplar to its matched images during encoding. For new concepts we extracted the highest similarity of a new concept to all old concepts from encoding.

DNN similarities as proxies for behavior in Study III

To determine the degree of format-specific processing during memory formation we constructed trial-specific measures indicating the distinctiveness of similarity judgements of task-relevant feature formats. For study I+II we capitalized on the behavioral ratings. Participants were instructed to rate the relation of both images on two dimensions, where the second image was rated relative to the first image. Thus, each image was considered a point in 2D space of which we computed its trial-specific Euclidean distance. The first image was the point of reference with coordinates [0,0] while the coordinates of the second image were the given ratings on each scale [rating1, rating2]. In study III we extracted our measure of distinctiveness from the pairwise similarities across DNN-derived layer predictions for all three images of a triple (target, similar choice, distractor choice) (see Supplementary Fig. 4). We took the similarity of the target image with the similar choice (r_{sim}) and the distractor choice (r_{dist}) and used their difference ($d = r_{sim} - r_{dist}$) as quantification for triplet (dis-) similarity. Both measures of trial-distinctiveness quantify the degree of difficulty in a given trial, e.g. small distances of animals indicate similar features, making the magnitude and direction of the given rating more difficult. Likewise, low triplet-distinctiveness indicates that both choice options (similar and distractor) are approximately equally similar to the target image, making the forced choice task more difficult.

References

1. Kunz, L., Deuker, L., Zhang, H. & Axmacher, N. Tracking Human Engrams Using Multivariate Analysis Techniques. In *Handbook of in Vivo Neural Plasticity Techniques* (Elsevier2018), Vol. 28, pp. 481–508.

2. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience* **2,** 4; 10.3389/neuro.06.004.2008 (2008).

3. Barry, D. N. & Maguire, E. A. Remote Memory and the Hippocampus: A Constructive Critique. *Trends in cognitive sciences* **23,** 128–142; 10.1016/j.tics.2018.11.005 (2019).

4. Cheng, S., Werning, M. & Suddendorf, T. Dissociating memory traces and scenario construction in mental time travel. *Neuroscience and biobehavioral reviews* **60,** 82–89; 10.1016/j.neubiorev.2015.11.011 (2016).

5. Gilboa, A. & Moscovitch, M. No consolidation without representation: Correspondence between neural and psychological representations in recent and remote memory. *Neuron* **109,** 2239–2255; 10.1016/j.neuron.2021.04.025 (2021).

6. Lifanov, J., Linde-Domingo, J. & Wimber, M. Feature-specific reaction times reveal a semanticisation of memories over time and with repeated remembering. *Nature communications* **12,** 3177; 10.1038/s41467-021-23288-5 (2021).

7. Brunec, I. K., Robin, J., Olsen, R. K., Moscovitch, M. & Barense, M. D. Integration and differentiation of hippocampal memory traces. *Neuroscience and biobehavioral reviews* **118,** 196–208; 10.1016/j.neubiorev.2020.07.024 (2020).

8. Heinen, R., Bierbrauer, A., Wolf, O. T. & Axmacher, N. Representational formats of human memory traces. *Brain structure & function*; 10.1007/s00429-023-02636-9 (2023).

9. Gifford, A. T., Dwivedi, K., Roig, G. & Cichy, R. M. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage* **264,** 119754; 10.1016/j.neuroimage.2022.119754 (2022).

10. Graumann, M., Ciuffi, C., Dwivedi, K., Roig, G. & Cichy, R. M. The spatiotemporal neural dynamics of object location representations in the human brain. *Nature human behaviour* **6,** 796–811; 10.1038/s41562-022-01302-0 (2022).

11. Clarke, A., Devereux, B. J. & Tyler, L. K. Oscillatory Dynamics of Perceptual to Conceptual Transformations in the Ventral Visual Pathway. *Journal of cognitive neuroscience* **30,** 1590–1605; 10.1162/jocn_a_01325 (2018).

12. Kietzmann, T. C. *et al.* Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences of the United States of America* **116,** 21854–21863; 10.1073/pnas.1905544116 (2019).

13. Kuzovkin, I. *et al.* Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications biology* **1,** 107; 10.1038/s42003-018-0110-y (2018).

14. Allen, E. J. *et al.* A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience* **25,** 116–126; 10.1038/s41593-021-00962-x (2022).

15. Güçlü, U. & van Gerven, M. A. J. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *The Journal of neuroscience : the official*

journal of the Society for Neuroscience **35,** 10005–10014; 10.1523/JNEUROSCI.5023-14.2015 (2015).

16. Lee, H. & Chen, J. Predicting memory from the network structure of naturalistic events. *Nature communications* **13,** 4235; 10.1038/s41467-022-31965-2 (2022).

17. Doerig, A. *et al.* The neuroconnectionist research programme. *Nature reviews. Neuroscience* **24,** 431–450; 10.1038/s41583-023-00705-w (2023).

18. Liu, J. *et al.* Transformative neural representations support long-term episodic memory. *Science advances* **7,** eabg9715; 10.1126/sciadv.abg9715 (2021).

19. Saxe, A., Nelli, S. & Summerfield, C. If deep learning is the answer, what is the question? *Nature reviews. Neuroscience* **22,** 55–67; 10.1038/s41583-020-00395-8 (2021).

20. Bellmund, J. L. S., Gärdenfors, P., Moser, E. I. & Doeller, C. F. Navigating cognition: Spatial codes for human thinking. *Science (New York, N.Y.)* **362**; 10.1126/science.aat6766 (2018).

21. Breedlove, J. L., St-Yves, G., Olman, C. A. & Naselaris, T. Generative Feedback Explains Distinct Brain Activity Codes for Seen and Mental Images. *Current biology : CB* **30,** 2211-2224.e6; 10.1016/j.cub.2020.04.014 (2020).

22. Linde-Domingo, J., Treder, M. S., Kerrén, C. & Wimber, M. Evidence that neural information flow is reversed between object perception and object reconstruction from memory. *Nature communications* **10,** 179; 10.1038/s41467-018-08080-2 (2019).

23. Davis, S. W. *et al.* Visual and Semantic Representations Predict Subsequent Memory in Perceptual and Conceptual Memory Tests. *Cerebral cortex (New York, N.Y. : 1991)* **31,** 974–992; 10.1093/cercor/bhaa269 (2021).

24. Liu, J. *et al.* Stable maintenance of multiple representational formats in human visual short-term memory. *Proceedings of the National Academy of Sciences of the United States of America* **117,** 32329–32339; 10.1073/pnas.2006752117 (2020).

25. Klinzing, J. G., Niethard, N. & Born, J. Mechanisms of systems memory consolidation during sleep. *Nature neuroscience* **22,** 1598–1610; 10.1038/s41593-019-0467-3 (2019).

26. Konkle, T., Brady, T. F., Alvarez, G. A. & Oliva, A. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General* **139,** 558–578; 10.1037/a0019165 (2010).

27. Winocur, G. & Moscovitch, M. Memory transformation and systems consolidation. *Journal of the International Neuropsychological Society : JINS* **17,** 766–780; 10.1017/S1355617711000683 (2011).

28. Dudai, Y., Karni, A. & Born, J. The Consolidation and Transformation of Memory. *Neuron* **88,** 20–32; 10.1016/j.neuron.2015.09.004 (2015).

29. Sols, I., DuBrow, S., Davachi, L. & Fuentemilla, L. Event Boundaries Trigger Rapid Memory Reinstatement of the Prior Events to Promote Their Representation in Long-Term Memory. *Current biology : CB* **27,** 3499-3504.e4; 10.1016/j.cub.2017.09.057 (2017).

30. Diekelmann, S. & Born, J. The memory function of sleep. *Nature reviews. Neuroscience* **11,** 114–126; 10.1038/nrn2762 (2010).

31. Lewis, P. A. & Bendor, D. How Targeted Memory Reactivation Promotes the Selective Strengthening of Memories in Sleep. *Current biology : CB* **29,** R906-R912; 10.1016/j.cub.2019.08.019 (2019).

32. Paller, K. A., Creery, J. D. & Schechtman, E. Memory and Sleep: How Sleep Cognition Can Change the Waking Mind for the Better. *Annual review of psychology* **72,** 123–150; 10.1146/annurev-psych-010419-050815 (2021).

33. Craik, F. I. M. Levels of processing: past, present. and future? *Memory (Hove, England)* **10,** 305–318; 10.1080/09658210244000135 (2002).

34. Lockhart, R. S. Levels of processing, transfer-appropriate processing, and the concept of robust encoding. *Memory (Hove, England)* **10,** 397–403; 10.1080/09658210244000225 (2002).

35. Niv, Y. Learning task-state representations. *Nature neuroscience* **22,** 1544–1553; 10.1038/s41593-019-0470-8 (2019).

36. Niv, Y. *et al.* Reinforcement learning in multidimensional environments relies on attention mechanisms. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **35,** 8145–8157; 10.1523/JNEUROSCI.2978-14.2015 (2015).

37. Radulescu, A., Shin, Y. S. & Niv, Y. Human Representation Learning. *Annual review of neuroscience* **44,** 253–273; 10.1146/annurev-neuro-092920-120559 (2021).

38. Bunzeck, N. & Düzel, E. Absolute coding of stimulus novelty in the human substantia nigra/VTA. *Neuron* **51,** 369–379; 10.1016/j.neuron.2006.06.021 (2006).

39. Frank, D. & Kafkas, A. Expectation-driven novelty effects in episodic memory. *Neurobiology of learning and memory* **183,** 107466; 10.1016/j.nlm.2021.107466 (2021).

40. Ranganath, C. & Rainer, G. Neural mechanisms for detecting and remembering novel events. *Nature reviews. Neuroscience* **4,** 193–202; 10.1038/nrn1052 (2003).

41. Restorff, H. von. *Über die Wirkung von Bereichsbildungen im Spurenfeld.* Berlin, Univ., Diss., 1933 (1933).

42. Bierbrauer, A., Fellner, M.-C., Heinen, R., Wolf, O. T. & Axmacher, N. The memory trace of a stressful episode. *Current biology : CB* **31,** 5204-5213.e8; 10.1016/j.cub.2021.09.044 (2021).

43. LaRocque, K. F. *et al.* Global similarity and pattern separation in the human medial temporal lobe predict subsequent memory. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **33,** 5466–5474; 10.1523/JNEUROSCI.4293-12.2013 (2013).

44. Sommer, V. R. & Sander, M. C. Contributions of representational distinctiveness and stability to memory performance and age differences. *Neuropsychology, development, and cognition. Section B, Aging, neuropsychology and cognition* **29,** 443–462; 10.1080/13825585.2021.2019184 (2022).

45. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60,** 84–90; 10.1145/3065386 (2017).

46. Cer, D. *et al.* Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, edited by E. Blanco & W. Lu (Association for Computational Linguistics, Stroudsburg, PA, USA, 2018), pp. 169–174.

47. Staresina, B. P., Gray, J. C. & Davachi, L. Event congruency enhances episodic memory encoding through semantic elaboration and relational binding. *Cerebral cortex (New York, N.Y. : 1991)* **19,** 1198–1207; 10.1093/cercor/bhn165 (2009).

48. van Kesteren, M. T. R., Ruiter, D. J., Fernández, G. & Henson, R. N. How schema and novelty augment memory formation. *Trends in neurosciences* **35,** 211–219; 10.1016/j.tins.2012.02.001 (2012).

49. McClelland, J. L. & Rumelhart, D. E. Distributed memory and the representation of general and specific information. *Journal of experimental psychology. General* **114,** 159–197; 10.1037//0096-3445.114.2.159 (1985).

50. Ofen, N. & Shing, Y. L. From perception to memory: changes in memory systems across the lifespan. *Neuroscience and biobehavioral reviews* **37,** 2258–2267; 10.1016/j.neubiorev.2013.04.006 (2013).

51. Sommer, V. R., Mount, L., Weigelt, S., Werkle-Bergner, M. & Sander, M. C. Memory specificity is linked to repetition effects in event-related potentials across the lifespan. *Developmental cognitive neuroscience* **48,** 100926; 10.1016/j.dcn.2021.100926 (2021).

52. Bartlett, F. C. *Remembering: A study in experimental and social psychology* (Cambridge University Press, New York, NY, US, 1932).

53. Spens, E. & Burgess, N. A generative model of memory construction and consolidation. *Nature human behaviour*; 10.1038/s41562-023-01799-z (2024).

54. Yonelinas, A. P., Ranganath, C., Ekstrom, A. D. & Wiltgen, B. J. A contextual binding theory of episodic memory: systems consolidation reconsidered. *Nature reviews. Neuroscience* **20,** 364–375; 10.1038/s41583-019-0150-4 (2019).

55. Xue, G. From remembering to reconstruction: The transformative neural representation of episodic memory. *Progress in neurobiology* **219,** 102351; 10.1016/j.pneurobio.2022.102351 (2022).

56. Barron, H. C., Auksztulewicz, R. & Friston, K. Prediction and memory: A predictive coding account. *Progress in neurobiology* **192,** 101821; 10.1016/j.pneurobio.2020.101821 (2020).

57. Greve, A., Cooper, E., Tibon, R. & Henson, R. N. Knowledge is power: Prior knowledge aids memory for both congruent and incongruent events, but in different ways. *Journal of experimental psychology. General* **148,** 325–341; 10.1037/xge0000498 (2019).

58. Fernández, R. S., Boccia, M. M. & Pedreira, M. E. The fate of memory: Reconsolidation and the case of Prediction Error. *Neuroscience and biobehavioral reviews* **68,** 423–441; 10.1016/j.neubiorev.2016.06.004 (2016).

59. Lisman, J. E. & Grace, A. A. The hippocampal-VTA loop: controlling the entry of information into long-term memory. *Neuron* **46,** 703–713; 10.1016/j.neuron.2005.05.002 (2005).

60. Schomaker, J. & Meeter, M. Short- and long-lasting consequences of novelty, deviance and surprise on brain and cognition. *Neuroscience and biobehavioral reviews* **55,** 268–279; 10.1016/j.neubiorev.2015.05.002 (2015).

61. Frank, D., Montemurro, M. A. & Montaldi, D. Pattern Separation Underpins Expectation-Modulated Memory. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **40,** 3455–3464; 10.1523/JNEUROSCI.2047-19.2020 (2020).

62. Garvert, M. M., Dolan, R. J. & Behrens, T. E. A map of abstract relational knowledge in the human hippocampal-entorhinal cortex. *eLife* **6**; 10.7554/eLife.17086 (2017).

63. Theves, S., Fernandez, G. & Doeller, C. F. The Hippocampus Encodes Distances in Multidimensional Feature Space. *Current biology : CB* **29,** 1226-1231.e3; 10.1016/j.cub.2019.02.035 (2019).

64. Buzsáki, G. & Moser, E. I. Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nature neuroscience* **16,** 130–138; 10.1038/nn.3304 (2013).

65. Rueckemann, J. W., Sosa, M., Giocomo, L. M. & Buffalo, E. A. The grid code for ordered experience. *Nature reviews. Neuroscience* **22,** 637–649; 10.1038/s41583-021-00499-9 (2021).

66. Tolman, E. C. Cognitive maps in rats and men. *Psychological review* **55,** 189–208; 10.1037/h0061626 (1948).

67. Cairney, S. A., Guttesen, A. Á. V., El Marj, N. & Staresina, B. P. Memory Consolidation Is Linked to Spindle-Mediated Information Processing during Sleep. *Current biology : CB* **28,** 948-954.e4; 10.1016/j.cub.2018.01.087 (2018).

68. Zhang, H., Fell, J. & Axmacher, N. Electrophysiological mechanisms of human memory consolidation. *Nature communications* **9,** 4103; 10.1038/s41467-018-06553-y (2018).

69. Wang, B. *et al.* Targeted Memory Reactivation during Sleep Elicits Neural Signals Related to Learning Content. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **39,** 6728–6736; 10.1523/JNEUROSCI.2798-18.2019 (2019).

70. Schreiner, T., Doeller, C. F., Jensen, O., Rasch, B. & Staudigl, T. Theta Phase-Coordinated Memory Reactivation Reoccurs in a Slow-Oscillatory Rhythm during NREM Sleep. *Cell reports* **25,** 296–301; 10.1016/j.celrep.2018.09.037 (2018).

71. Fuentemilla, L. *et al.* Hippocampus-dependent strengthening of targeted memories via reactivation during sleep in humans. *Current biology : CB* **23,** 1769–1775; 10.1016/j.cub.2013.07.006 (2013).

72. Himmer, L., Schönauer, M., Heib, D. P. J., Schabus, M. & Gais, S. Rehearsal initiates systems memory consolidation, sleep makes it last. *Science advances* **5,** eaav1695; 10.1126/sciadv.aav1695 (2019).

73. Lewis, P. A. & Durrant, S. J. Overlapping memory replay during sleep builds cognitive schemata. *Trends in cognitive sciences* **15,** 343–351; 10.1016/j.tics.2011.06.004 (2011).

74. Liu, Y., Dolan, R. J., Kurth-Nelson, Z. & Behrens, T. E. J. Human Replay Spontaneously Reorganizes Experience. *Cell* **178,** 640-652.e14; 10.1016/j.cell.2019.06.012 (2019).

75. Naspi, L., Hoffman, P., Devereux, B. & Morcom, A. M. Perceptual and Semantic Representations at Encoding Contribute to True and False Recognition of Objects. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **41,** 8375–8389; 10.1523/jneurosci.0677-21.2021 (2021).

76. Deuker, L. *et al.* Memory consolidation by replay of stimulus-specific neural activity. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **33,** 19373–19383; 10.1523/JNEUROSCI.0414-13.2013 (2013).

77. Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D. Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron* **93,** 480–490; 10.1016/j.neuron.2016.12.041 (2017).

78. Theves, S., Fernández, G. & Doeller, C. F. The Hippocampus Maps Concept Space, Not Feature Space. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **40,** 7318–7325; 10.1523/JNEUROSCI.0494-20.2020 (2020).

79. Donoghue, T. *et al.* Single neurons in the human medial temporal lobe flexibly shift representations across spatial and memory tasks. *bioRxiv : the preprint server for biology*; 10.1101/2023.02.22.529437 (2023).

80. Strange, B. A., Witter, M. P., Lein, E. S. & Moser, E. I. Functional organization of the hippocampal longitudinal axis. *Nature reviews. Neuroscience* **15,** 655–669; 10.1038/nrn3785 (2014).

81. Collin, S. H. P., Milivojevic, B. & Doeller, C. F. Memory hierarchies map onto the hippocampal long axis in humans. *Nature neuroscience* **18,** 1562–1564; 10.1038/nn.4138 (2015).

82. Przeździk, I., Faber, M., Fernández, G., Beckmann, C. F. & Haak, K. V. The functional organisation of the hippocampus along its long axis is gradual and predicts recollection. *Cortex; a journal devoted to the study of the nervous system and behavior* **119,** 324–335; 10.1016/j.cortex.2019.04.015 (2019).

83. Constantinescu, A. O., O'Reilly, J. X. & Behrens, T. E. J. Organizing conceptual knowledge in humans with a gridlike code. *Science (New York, N.Y.)* **352,** 1464–1468; 10.1126/science.aaf0941 (2016).

84. Viganò, S., Rubino, V., Buiatti, M. & Piazza, M. The neural representation of absolute direction during mental navigation in conceptual spaces. *Communications biology* **4,** 1294; 10.1038/s42003-021-02806-7 (2021).

85. Liang, Z. *et al.* Social navigation: distance and grid-like codes support navigation of abstract social space in human brain. *eLife,* 12:RP89025; 10.7554/eLife.89025.1 (2023).

86. Huang, S. *et al.* Hippocampal functions modulate transfer-appropriate cortical representations supporting subsequent memory. *The Journal of neuroscience : the official journal of the Society for Neuroscience*; 10.1523/JNEUROSCI.1135-23.2023 (2023).

87. Jozwik, K. M., Kietzmann, T. C., Cichy, R. M., Kriegeskorte, N. & Mur, M. Deep Neural Networks and Visuo-Semantic Models Explain Complementary Components of Human Ventral-Stream Representational Dynamics. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **43,** 1731–1741; 10.1523/JNEUROSCI.1424-22.2022 (2023).

88. King, M. L., Groen, I. I. A., Steel, A., Kravitz, D. J. & Baker, C. I. Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage* **197,** 368–382; 10.1016/j.neuroimage.2019.04.079 (2019).

89. Leeuw, J. R. de. jsPsych: a JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods* **47,** 1–12; 10.3758/s13428-014-0458-y (2015).

90. Peirce, J. *et al.* PsychoPy2: Experiments in behavior made easy. *Behavior research methods* **51,** 195–203; 10.3758/s13428-018-01193-y (2019).

91. Hebart, M. N. *et al.* THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one* **14,** e0223792; 10.1371/journal.pone.0223792 (2019).

92. Whitmore, N. W., Bassard, A. M. & Paller, K. A. Targeted memory reactivation of face-name learning depends on ample and undisturbed slow-wave sleep. *NPJ science of learning* **7,** 1; 10.1038/s41539-021-00119-2 (2022).

93. Oudiette, D. & Paller, K. A. Upgrading the sleeping brain with targeted memory reactivation. *Trends in cognitive sciences* **17,** 142–149; 10.1016/j.tics.2013.01.006 (2013).

94. Rudoy, J. D., Voss, J. L., Westerberg, C. E. & Paller, K. A. Strengthening individual memories by reactivating them during sleep. *Science (New York, N.Y.)* **326,** 1079; 10.1126/science.1179013 (2009).

95. van Dongen, E. V. *et al.* Memory stabilization with targeted reactivation during human slow-wave sleep. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 10575–10580; 10.1073/pnas.1201072109 (2012).

96. Gais, S., Plihal, W., Wagner, U. & Born, J. Early sleep triggers memory for early visual discrimination skills. *Nature neuroscience* **3,** 1335–1339; 10.1038/81881 (2000).

97. Plihal, W. & Born, J. Effects of early and late nocturnal sleep on priming and spatial memory. *Psychophysiology* **36,** 571–582 (1999).

98. Yordanova, J., Kolev, V., Wagner, U. & Verleger, R. Differential associations of early- and late-night sleep with functional brain states promoting insight to abstract task regularity. *PloS one* **5,** e9442; 10.1371/journal.pone.0009442 (2010).

99. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference* (SciPy2010), pp. 92–96.

100. Jia, Y. *et al.* Caffe. In *Proceedings of the 22nd ACM international conference on Multimedia,* edited by K. A. Hua, *et al.* (ACM, New York, NY, USA, 2014), pp. 675–678.

101. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, 14.03.2016.

102. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* **17,** 261–272; 10.1038/s41592-019-0686-2 (2020).

Supplementary Materials

Supplementary Figure 1.
Mixed blocks reflect perceptual and conceptual formats on single dimension levels

Supplementary Figure 2.
Representational distance affects subsequent memory regardless of perceptual/conceptual encoding space and cueing

Supplementary Figure 3.
Layer-wise and averaged representational distances in Deep Neural Networks reflect behavioral similarity judgements

Supplementary Figure 4.
Layer-wise correspondence of representational distances in DNNs to binary ratings in Study III

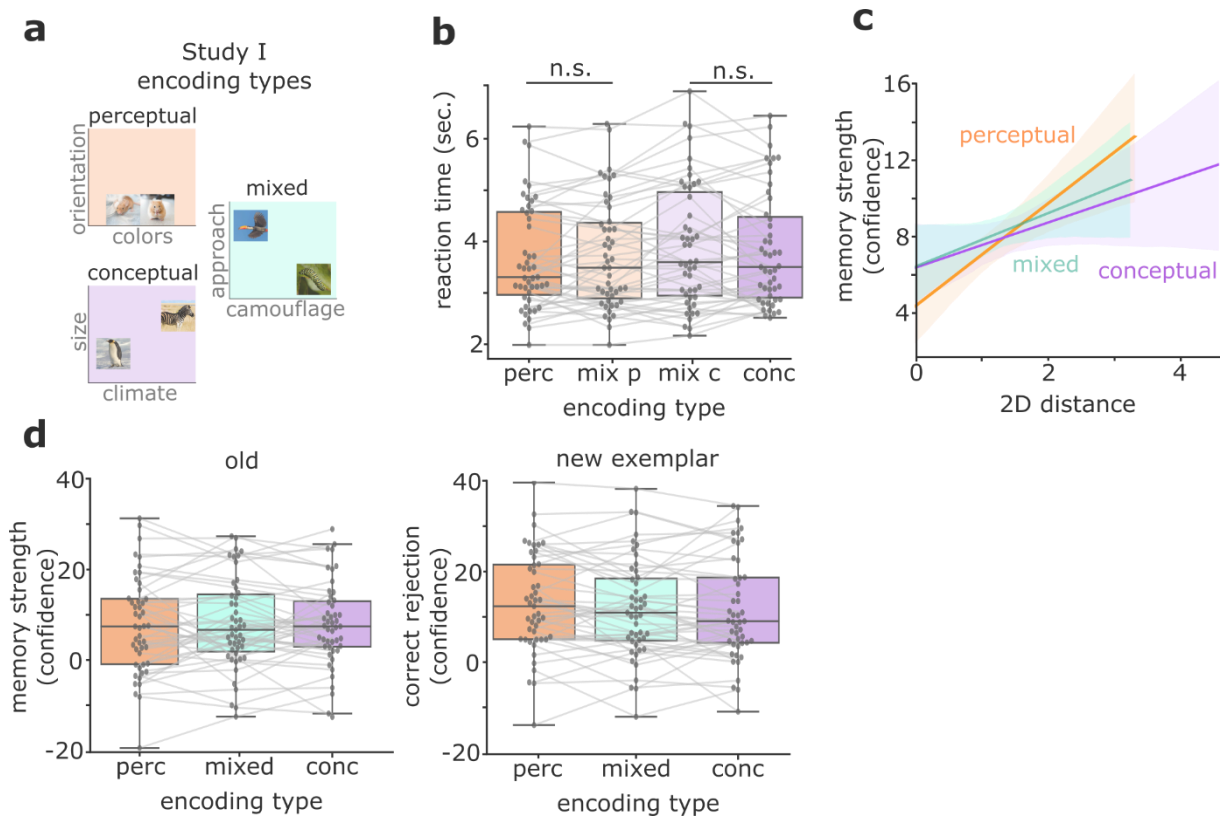
Supplementary Figure 5.
Effects of DNN-based representational similarities on memory strength during the general and the specific memory task

Supplementary Figure 6.
Different formats contribute to memory strength and correct rejection performance depending on task instructions at retrieval

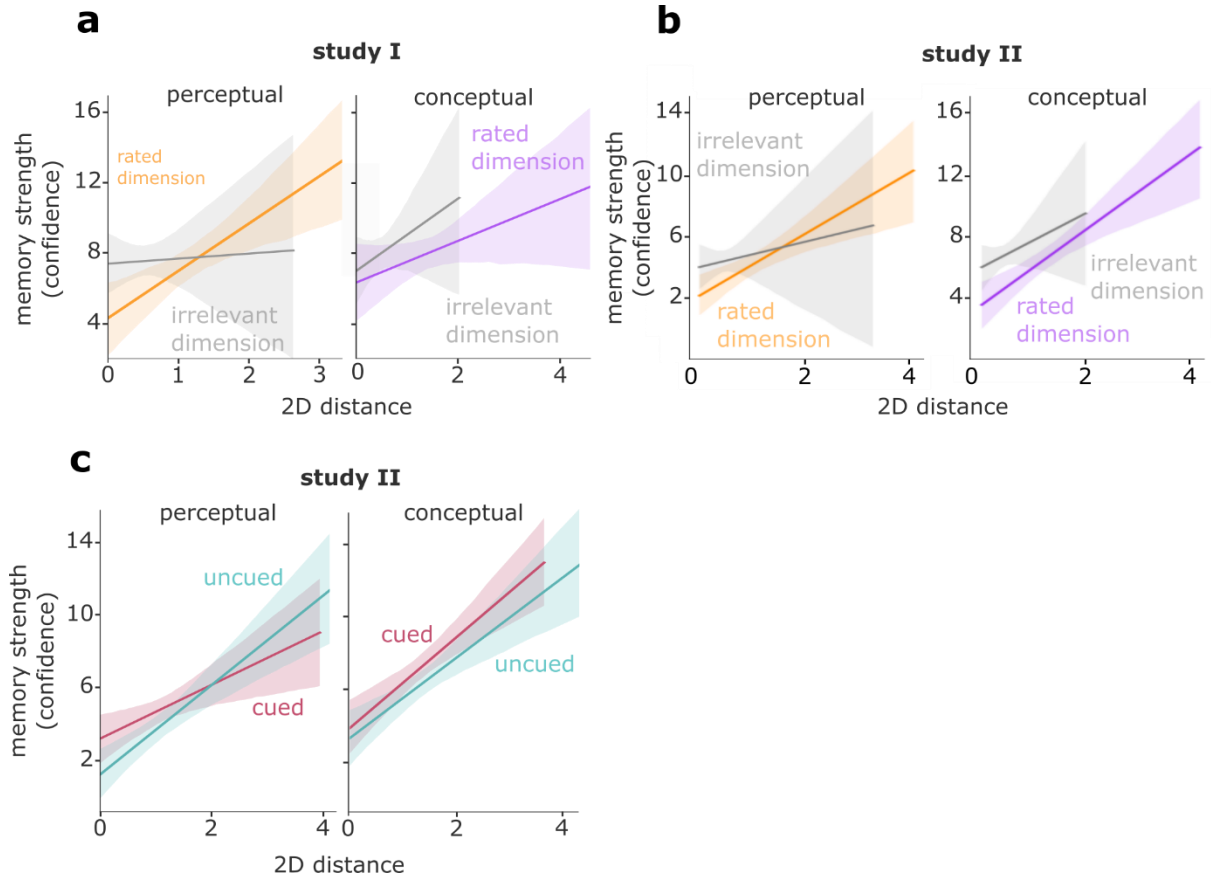
Supplementary Figure 7.
Conceptual cueing specifically influences semantic memory

Supplementary Table 1.
Linear mixed models and their beta weights from main results

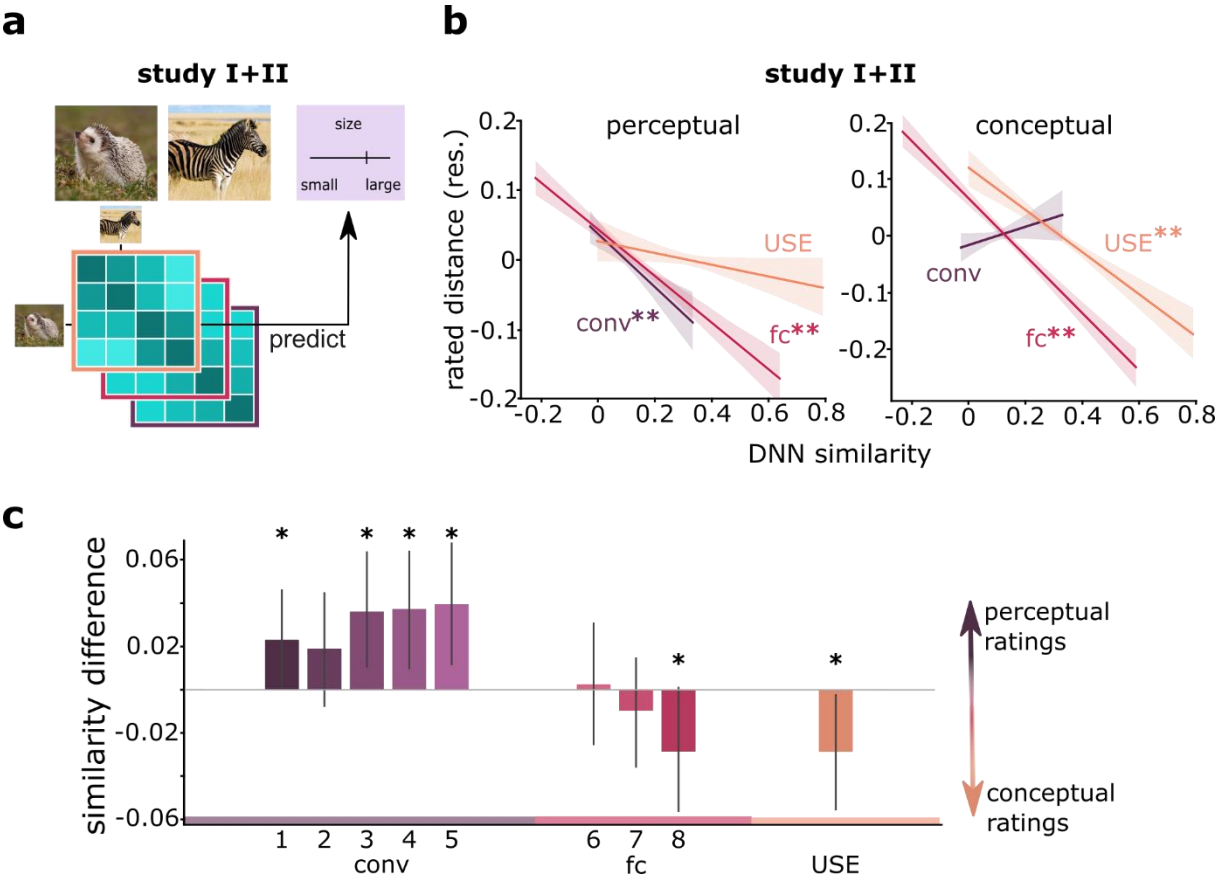
Supplementary Table 2.
Linear mixed models and their beta weights from Supplementary analyses



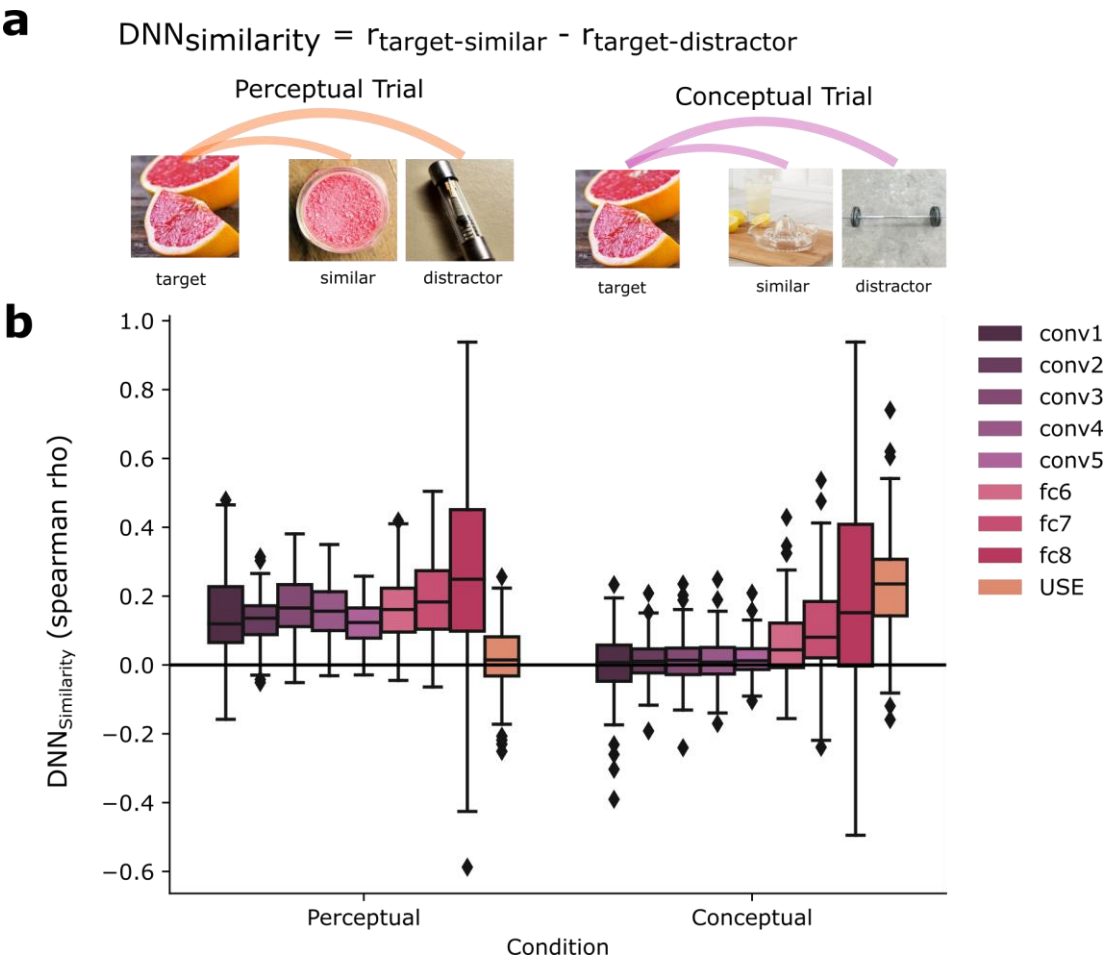
Supplementary Figure 1. Mixed blocks reflect perceptual and conceptual formats on single dimension levels: a) In Study I, we presented perceptual blocks (two perceptual dimensions), conceptual blocks (two conceptual dimensions), and mixed blocks (one perceptual and one conceptual dimension). We investigated whether ratings in mixed blocks would differ from ratings in congruent (i.e., purely perceptual, and purely conceptual) blocks. b) We used linear mixed model to predict reaction times using encoding type (perceptual/conceptual/mixed) as a predictor, including a random participant intercept. Reaction times during perceptual ratings in mixed blocks ($z = -1.65$, $\chi^2_{(1)} = 2.72$, $p = 0.098$; likelihood ratio test) and during conceptual ratings in mixed blocks ($z = 1.38$, $\chi^2_{(1)} = 1.91$, $p = 0.167$; likelihood ratio test) did not differ from their congruent counterpart. c) We then assessed whether the influence of rated distances on memory strength differs between encoding spaces. Using a linear mixed model including the interaction of encoding space (conceptual, perceptual, mixed) and 2D distance as predictors, we found no interaction of 2D distances and encoding space (perceptual vs conceptual: $z = 1.05$, mixed vs conceptual: $z = 1.04$, mixed vs perceptual: $z = 1.38$, $\chi^2_{(2)} = 2.89$, $p = 0.235$; likelihood ratio test). d) Memory strength (left; perceptual versus conceptual: $z = 0.49$, mixed vs conceptual: $z = -0.11$, mixed vs perceptual: $z = 0.64$, $\chi^2_{(1)} = 0.42$, $p = 0.810$; likelihood ratio test) and correct rejection of new exemplars (right; conceptual vs perceptual: $z = -1.74$, mixed vs perceptual: $z = -1.61$, mixed vs conceptual: $z = -0.10$, $\chi^2_{(1)} = 3.78$, $p = 0.151$; likelihood ratio test) did not differ between blocks. Based on these results, we decided to concatenate perceptual dimension ratings from mixed and perceptual blocks and conceptual ratings from mixed and conceptual blocks for further analysis of DNN-based similarities in Study I. Data are visualized after removing participant-wise estimated random effects. 95% confidence intervals, center line indicates median, box limits indicate upper and lower quartiles, whiskers indicate 1.5 interquartile range, individual points represent participant averages. n.s., not significant



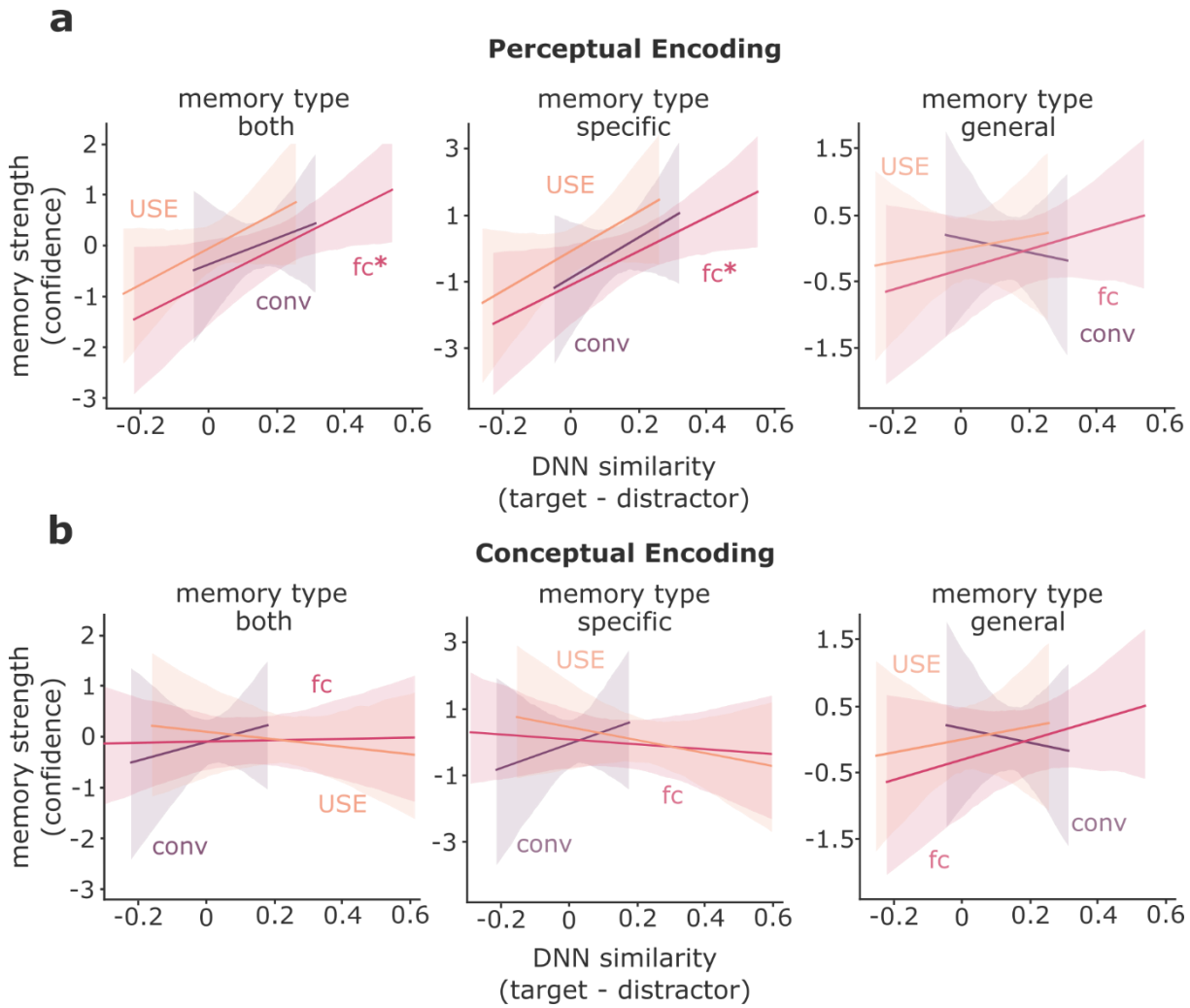
Supplementary Figure 2. Representational distance affects subsequent memory regardless of perceptual/conceptual encoding space and cueing. a) Study I: Encoding space did not affect the impact of representational distance on subsequent memory (interaction space x rated distance: z = 1.56, $\chi^2_{(2)} = 2.89$, p = 0.235). b) Same for Study II (interaction space x rated distance: z = -0.63, $\chi^2_{(2)} = 0.40$, p = 0.820). c) Study II: Cueing did not affect the impact of representational distance neither for perceptual (interaction cueing x rated distance [perceptual trials only]: z = 1.14, $\chi^2_{(2)} = 1.30$, p = 0.52) nor for conceptual encoding spaces (interaction cueing x rated distance [conceptual trials only]: z = -0.11, $\chi^2_{(2)} = 0.01$, p = 0.99). Data are visualized after removing participant-wise estimated random effects. 95% confidence intervals.



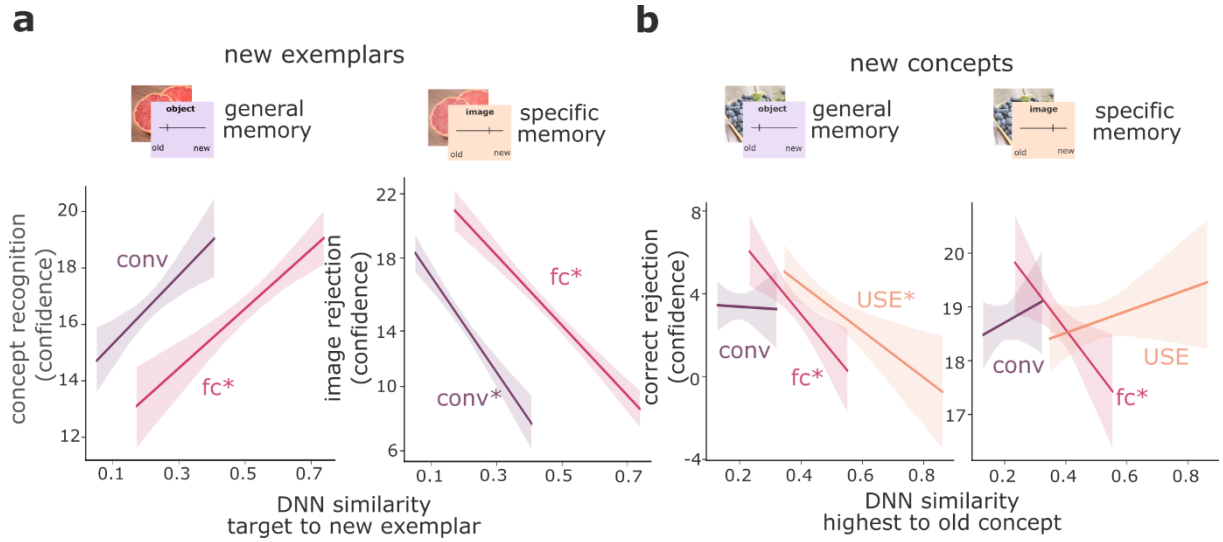
Supplementary Figure 3. Layer-wise and averaged representational distances in Deep Neural Networks reflect behavioral similarity judgements. a) We tested whether similarity judgements in the respective task-relevant perceptual and conceptual spaces corresponded to representational distances in specific DNN layers. To this end, we computed the similarity of each image pair in each network layer of a convolutional DNN and in the universal sentence encoder (USE) and used these similarities as predictors in a linear mixed model on rated distances of Study I and II. b) Using the layer averages (see Methods for more details) of convolutional (conv) and fully connected (fc) layers as well as the USE model similarities, we found that low-level perceptual and higher-order perceptual similarities predict rated distances during perceptual encoding. Reversely, conceptual ratings can be predicted by higher-order perceptual and semantic similarities (see Results for statistics). c) In addition, we analyzed layer-wise similarities using paired t-tests (one-sided), FDR corrected for multiple comparisons. Effect sizes are measured by Cohen's d. Indeed, representational distances in cDNN layers conv1 (t_{101} = 1.86, p = 0.049, d = 0.18, perceptual: M = 0.024, SD = 0.1, conceptual: M = 0.001, SD = 0.08), conv3 (t_{101} = 2.58, p = 0.017, d = 0.25, perceptual: M = 0.074, SD = 0.1, conceptual: M = 0.038, SD = 0.09), conv4 (t_{101} = 2.69, p = 0.017, d = 0.26, perceptual: M = 0.07, SD = 0.09, conceptual: M = 0.03, SD = 0.08) and conv5 (t_{101} = 2.88, p = 0.017, d = 0.28, perceptual: M = 0.08, SD = 09, conceptual: M = 0.04, SD = 0.08) predicted perceptual ratings better than conceptual ratings, while the reverse was true for fully-connected layer 8 (t_{101} = -1.93, p = 0.049, d = 0.19, perceptual: M = 0.10, SD = 0.09, conceptual: M = 0.13, SD = 0.10) and USE distances (t_{101} = -2.15, p = 0.037, d = 0.21, perceptual: M = 0.04, SD = 0.10, conceptual: M = 0.07, SD = 0.09) reflects the mean over all Spearman rho correlations of each DNN layer across all participant ratings). Data are visualized after removing participant-wise estimated random effects. 95% confidence intervals, error bars indicate SEM. *, p<0.05, **, p<0.01



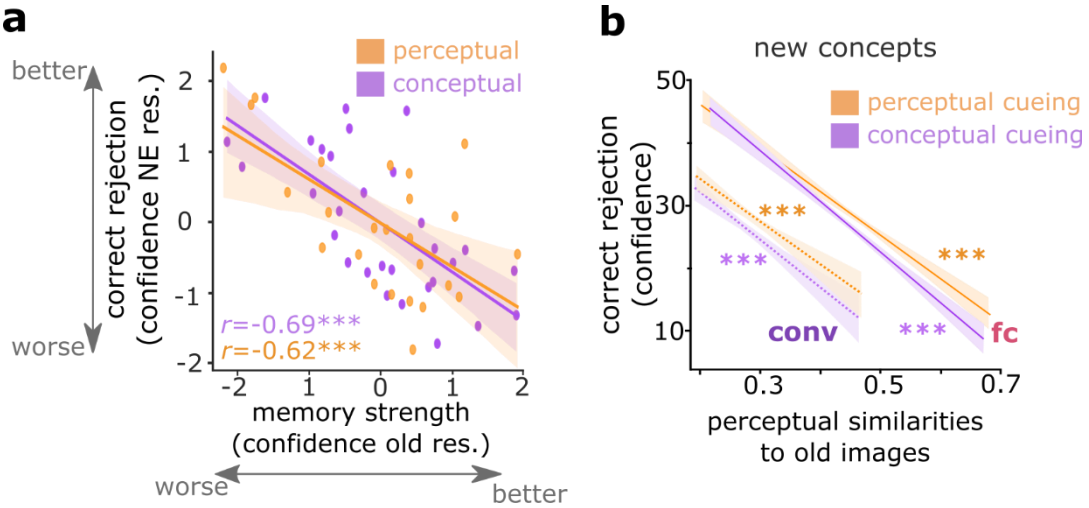
Supplementary Figure 4. Layer-wise correspondence of representational distances in DNNs to binary ratings in Study III. a) Exemplary target image with perceptual and conceptual choice options (similar choice, distractor choice). Lines indicate pairwise image similarities used to compute a triplet similarity score ($d_{similarity}$). b) Distribution of $d_{similarity}$ scores across cDNN layers for conceptual and perceptual trials. Values larger than zero indicate higher DNN-based representational similarities of targets with the similar image as compared to the distractor image. For each layer and condition, boxplots depict the distribution of $d_{similarity}$ scores across triplets. For perceptual triplets, $d_{similarity}$ scores are significantly larger than zero for all cDNN layers, but not for the USE. For conceptual triplets, $d_{similarity}$ scores are significantly larger than zero from conv2 until and including the USE model. Center line indicates median, box limits indicate upper and lower quartiles, whiskers indicate 1.5 interquartile range, points represent outliers.



Supplementary Figure 5. Effects of DNN-based representational similarities on memory strength during the general and the specific memory task. a) Prediction of memory strength by DNN similarities during perceptual choices. Left: Collapsed across memory types, representational distances in fully-connected layers but not convolutional layers (conv) and not semantic distances (USE-based) predict memory strength (see Results section). Middle: Memory strength in the specific retrieval task is predicted by fully-connected (: z = 2.258, $\chi^2_{(1)} = 5.09$, p = 0.024) but not convolutional or semantic (conv: z = 1.138, $\chi^2_{(1)} = 1.29$, p = 0.255; USE: z = 1.567, $\chi^2_{(1)} = 2.45$, p = 0.117; likelihood ratio test) DNN similarities during perceptual encoding. Right: Memory in the general memory test is independent of DNN similarities during perceptual encoding (conv: z = -0.293, $\chi^2_{(1)} = 0.09$, p = 0.769; fc: z = 0.977, $\chi^2_{(1)} = 0.95$, p = 0.328; USE: z = 0.373, $\chi^2_{(1)} = 0.14$, p = 0.709). b) Prediction of memory strength by DNN similarities during conceptual encoding. Left: Collapsed across memory types, DNN-predicted similarities do not predict subsequent memory (see Results section). Middle: Memory strength during specific retrieval cannot be predicted by DNN similarities (conv: z = 0.667, $\chi^2_{(1)} = 0.45$, p = 0.504; fc: z = -0.404, $\chi^2_{(1)} = 0.16$, p = 0.686; USE: z = -0.807, $\chi^2_{(1)} = 0.65$, p = 0.419; likelihood ratio test). Right: Memory strength during general retrieval cannot be predicted by DNN similarities. For studies I and II, where task-demands did not differ between general and specific retrieval, we found no effect of DNN similarities on memory strengths (Study I+II: conv: z = 0.49, $\chi^2_{(1)} = 1.34$, p = 0.247; fc: z = 0.94, $\chi^2_{(1)} = 1.00$, p = 0.317; USE: z = 0.87, $\chi^2_{(1)} = 1.90$, p = 0.167; likelihood ratio test). Data are visualized after removing participant-wise estimated random effects. 95% confidence intervals. *p<0.05



Supplementary Figure 6. Different formats contribute to memory strength and correct rejection performance depending on task instructions at retrieval. a) We computed the similarity of each new exemplar image to the old image of the same category and extracted the similarity value as the DNN similarity (for each DNN layer). Using a linear mixed model including both the main effects of DNN layers and task, we found significant effects of retrieval instruction with distances in low-level perceptual space ($z = -2.22$, $\chi^2_{(1)} = 13.30$, $p = 0.0002$; likelihood ratio test) and higher-order perceptual space ($z = -7.89$, $\chi^2_{(1)} = 70.34$, $p<0.0001$; likelihood ratio test) indicating that representational distances between new and previously presented items exerted different effects depending on retrieval tasks. Correct rejection performance for new exemplars in the specific retrieval task depended on similarities of both low-level perceptual ($z = -2.51$, $\chi^2_{(1)} = 6.34$, $p = 0.011$; likelihood ratio test) and higher-order perceptual ($z = -7.16$, $\chi^2_{(1)} = 50.97$, $p<0.0001$; likelihood ratio test) formats, while memory strength (confidence) for old concepts during general retrieval depended on similarity of higher-order ($z = 4.40$, $\chi^2_{(1)} = 19.37$, $p<0.0001$; likelihood ratio test) but not low-level perceptual formats ($z = 0.51$, $\chi^2_{(1)} = 0.26$, $p = 0.6090$; likelihood ratio test). (Note that semantic formats could not be considered in this analysis as new exemplars always belonged to the same concept.) b) For new concepts we calculated the semantic similarity to all encoding concepts and extracted the highest value (highest similarity of new concept – all old images/concepts): Using a linear mixed model including DNN similarities as predictors, we found that higher-order perceptual ($z = -2.78$, $\chi^2_{(1)} = 7.73$, $p = 0.005$; likelihood ratio test) but not low-level perceptual ($z = 1.39$, $\chi^2_{(1)} = 1.95$, $p = 0.162$; likelihood ratio test) or USE ($z = 0.50$, $\chi^2_{(1)} = 0.25$, $p = 0.615$; likelihood ratio test) similarities predicted performance during specific retrieval. During general retrieval, both higher-order perceptual ($z = -4.31$, $\chi^2_{(1)} = 18.57$, $p < 0.0001$; likelihood ratio test) and semantic similarities ($z = -3.96$, $\chi^2_{(1)} = 15.65$, $p < 0.0001$; likelihood ratio test) but not low-level perceptual similarities ($z = 1.14$, $\chi^2_{(1)} = 1.32$, $p = 0.250$; likelihood ratio test) predicted rejection of new concepts. Data are visualized after removing participant-wise estimated random effects. 95% confidence intervals. *, $p<0.05$



Supplementary Figure 7. Conceptual cueing specifically influences semantic memory. a) We tested whether participants who showed higher recognition memory performance also showed a decrease in rejection performance of new exemplars because they generally used a more liberal criterion during memory decisions. To this end, we calculated Pearson's correlations on participants' mixed model residuals. We found that participants in both conceptual (r = -0.69, p<0.001) and perceptual (r = -0.62, p<0.001) cueing conditions show this effect, i.e., those who performed better in identifying old images made more errors for new exemplars. b) Since conceptual cueing impaired the rejection of new concepts that were semantically highly similar to old concepts, we tested whether cueing affected the rejection of perceptually similar new concepts as well. To this end – analogously to our analysis on semantic similarities (Fig 5d) – we computed two linear mixed models (one for perceptual cueing and one for conceptual cueing) using perceptual similarities from the cDNN as predictors (highest similarity new concept – old images). Perceptual features predicted correct rejection confidence in both perceptual (conv: z = -6.64, $\chi^2_{(1)}$ = 43.84, p < 0.001; fc: z = -13.71, $\chi^2_{(1)}$ = 183.88, p < 0.001) and conceptual cueing conditions (conv: z = -7.21, $\chi^2_{(1)}$ = 51.69, p < 0.001; fc: z = -15.46, $\chi^2_{(1)}$ = 233.53, p < 0.001). Thus, low-level perceptual (conv) and higher-order perceptual (fc) similarities impaired rejection performance following both perceptual and conceptual cueing. Data are visualized after removing participant-wise estimated random effects. 95% confidence intervals. ***, p<0.0

Supplementary Table 1. Linear mixed models and their beta weights from main results. Corresponding Beta weights to the z values reported in the main results. All models include a random participant intercept (1|participant).

Linear mixed model		β	*p*
Reaction times (SI-III)			
Reaction times ~ rated distance 2D (**I**)		-277.82	<0.001
Reaction times ~ rated distance 2D (**II**)		-148.62	<0.001
*Reaction times ~ conv + fc + USE (**III**: perceptual encoding)*			
	Conv	-0.98	<0.001
	Fc	-0.14	0.196
	USE	0.07	0.639
*Reaction times ~ conv + fc + USE (**III**: conceptual encoding)*			
	Conv	0.11	0.621
	Fc	-0.39	<0.001
	USE	-0.28	0.007
Confidence for old images by encoding space (SI+SII)			
Memory confidence (old images) ~ encoding space (**I**)		-0.22	0.716
Memory confidence (uncued old images) ~ encoding space (**II**)		0.49	0.875
Confidence for old images by 2D distance (SI+II)			
Memory confidence ~ relevant distance (**I**)		1.76	<0.001
Memory confidence ~ relevant distance (**II**)		2.10	<0.001
Memory confidence ~ irrelevant distance + relevant distance (**I**)		0.72	0.431

Linear mixed model		β	*p*
Confidence for old images by encoding space using DNNs (SIII)			
Memory confidence ~ conv + fc + USE (perceptual encoding)			
	Conv	2.91	0.460
	Fc	3.91	0.022
	USE	3.71	0.151
Memory confidence ~ conv + fc + USE (conceptual encoding)			
	Conv	2.32	0.597
	Fc	0.16	0.913
	USE	-0.76	0.632
Confidence for old images by encoding space (SIII)			
*Memory confidence ~ encoding type + memory type + encoding space*memory type*			
	encoding space*memory type	1.64	0.007
Memory confidence ~ encoding space (specific memory)		0.66	0.198
Memory confidence ~ encoding space (general memory)		-0.98	0.002
Confidence for new exemplars by encoding space (SIII)			
*Confidence ~ encoding type + memory type + encoding space *memory type*			
	encoding space*memory type	1.52	0.011
Rejection confidence ~ encoding space (specific memory)		-0.17	0.680

Supp Table 1. continued

Memory confidence ~ irrelevant distance + relevant distance (**II**)	0.95	0.332	*Memory confidence ~ encoding space (general memory)*	-1.73	<0.001
Confidence for old images by 2D distance depending on temporal order			**Rejection confidence for new exemplars by 2D distance (SI+II)**		
Memory confidence (first image) ~ relevant distance (**I**)	0.98	0.100	*Rejection confidence ~ relevant distance* (**I**)	-0.83	0.129
Memory confidence (second image) ~ relevant distance (**I**)	2.51	<0.001	*Rejection confidence ~ relevant distance* (**II**)	-0.36	0.434
Memory confidence (first image) ~ relevant distance (**II**)	2.37	<0.001	**TMR effects on rejection confidence for new exemplars and new concepts (SII)**		
Memory confidence (second image) ~ relevant distance (**II**)	2.27	<0.001	*Rejection confidence (new exemplars) ~ encoding space + cueing*		
TMR effects on confidence for old images (SII)			Encoding space	1.20	0.684
Memory confidence (old images) ~ encoding space + cueing			Cueing	0.024	0.969
Encoding space	1.66	0.552	*Rejection confidence (new exemplars) ~ encoding space + cueing + encoding space*cueing*		
Cueing	1.04	0.033	Encoding space*cueing	4.62	<0.001
*Memory confidence (old images) ~ encoding space + cueing + encoding space*cueing*			*Rejection confidence (new exemplars) ~ cueing (cued conceptual encoding)*	2.12	0.017
Encoding space*Cueing	4.31	<0.001	*Rejection confidence (new exemplars) ~ cueing (cued perceptual encoding)*	-2.49	0.006
Memory confidence (old images) ~ cueing (cued conceptual encoding)	3.00	<0.001	*Rejection confidence (new concepts) ~ cueing (cued conceptual encoding)*	-12.34	0.008
Memory confidence (old images) ~ cueing (cued perceptual encoding)	1.30	0.060	*Rejection confidence (new concepts) ~ cueing (cued perceptual encoding)*	-7.11	0.109
			Rejection confidence (new exemplars, uncued) ~ encoding space	1.10	0.726

Supplementary Table 2. Linear mixed models and their beta weights from Supplementary Analyses. Corresponding Beta weights to the z values reported in the Supplementary Analyses. All models include a random participant intercept (1|participant).

Linear mixed model	β	*p*	Linear mixed model	β	*p*
Supplementary Fig. 1: Reaction times by distance in mixed vs congruent blocks (SI)					
Reaction times (perceptual trials) ~ rated distance + congruency (mixed/congruent)			*Reaction times (conceptual trials) ~ rated distance + congruency* (mixed/congruent)		
Congruency	67.18	0.099	Congruency	59.35	0.167
Supplementary Fig. 1: Confidence for old images by space (SI)			**Supplementary Fig. 1: Rejection confidence for new exemplars by space (SI)**		
Memory confidence ~ space			*Rejection confidence ~ space*		
Mixed (vs conceptual)	-0.17	0.777	Mixed (vs conceptual)	0.08	0.915
Perceptual (vs conceptual)	0.22	0.716	Perceptual (vs conceptual)	1.41	0.082
Perceptual (vs mixed)	0.40	0.518	Perceptual (vs mixed)	1.32	0.102
Supplementary Fig. 1: Confidence for old images by 2D distance and space (SI)			**Supplementary Fig. 1: Confidence for old images by 2D distance and cueing (SI+II)**		
*Memory confidence ~ 2D distance + space + 2D distance*space*			*Memory confidence ~ 2D distance + 2D distance*cueing (perceptual cueing)*		
2D*perceptual (vs conceptual)	1.64	0.118	2D*cueing	0.89	0.254
2D*mixed (vs conceptual)	0.16	0.874	*Memory confidence ~ 2D distance + 2D distance*cueing (conceptual cueing)*		
2D*mixed (vs perceptual)	1.47	0.167	2D*cueing	-0.08	0.911
Supplementary Fig. 3: Rated distances predicted by DNN layer (SI+II)					
Rated distance ~ conv + fc + USE (perceptual encoding)			*Rated distance ~ conv + fc + USE* (conceptual encoding)		
Conv	-0.47	<0.001	Conv	0.20	0.078
Fc	-0.44	<0.001	Fc	-0.67	<0.001
USE	-0.09	0.060	USE	-0.40	<0.001

Supplementary Table 2. continued

Supplementary Fig. 5: Confidence by space and retrieval task (SIII)						
Memory confidence (old images) ~ conv + fc + USE (perceptual encoding|specific memory)				*Memory confidence (old images) ~ conv + fc + USE* (conceptual encoding|specific memory)		
	Conv	7.11	0.255		Conv	4.69	0.504
	Fc	6.11	0.024		Fc	-0.95	0.686
	USE	6.41	0.117		USE	-2.04	0.420
Memory confidence (old images) ~ conv + fc + USE (perceptual encoding|general memory)				*Memory confidence (old images) ~ conv + fc + USE* (conceptual encoding|general memory)		
	Conv	-1.21	0.770		Conv	-0.03	0.993
	Fc	1.74	0.329		Fc	1.28	0.347
	USE	1.00	0.709		USE	0.52	0.718
Supplementary Fig. 6: Confidence for new exemplars and new concepts by encoding space and retrieval task (SIII)						
Memory confidence (new exemplars) ~ conv + fc (general memory)				*Rejection confidence (new exemplars) ~ conv + fc* (specific memory)		
	Conv	1.63	0.609		Conv	-9.08	0.012
	Fc	7.90	<0.001		Fc	-14.51	<0.001
Rejection confidence (new concepts) ~ conv + fc + USE (general memory)				*Rejection confidence (new concepts) ~ conv + fc + USE* (specific memory)		
	Conv	6.67	0.250		Conv	4.61	0.163
	Fc	-17.92	<0.001		Fc	-6.57	0.005
	USE	-10.96	<0.001		USE	0.79	0.616
Supplementary Fig. 7: Rejection confidence of new concepts depending on highest perceptual similarity (new concept – all old images) and cueing (SII)						
Rejection confidence (perceptual cueing) ~ conv + fc				*Rejection confidence (conceptual cueing) ~ conv + fc*		
	Conv	-186.33	<0.001		Conv	-239.82	<0.001
	Fc	-81.31	<0.001		Fc	-89.22	<0.001

1217