Brain Structure and Function (2024) 229:513–529
https://doi.org/10.1007/s00429-023-02636-9

REVIEW

Representational formats of human memory traces

Rebekka Heinen[1] · Anne Bierbrauer[1,2] · Oliver T. Wolf[3] · Nikolai Axmacher[1]

Received: 6 December 2022 / Accepted: 28 March 2023 / Published online: 6 April 2023

Abstract

Neural representations are internal brain states that constitute the brain's model of the external world or some of its features. In the presence of sensory input, a representation may reflect various properties of this input. When perceptual information is no longer available, the brain can still activate representations of previously experienced episodes due to the formation of memory traces. In this review, we aim at characterizing the nature of neural memory representations and how they can be assessed with cognitive neuroscience methods, mainly focusing on neuroimaging. We discuss how multivariate analysis techniques such as representational similarity analysis (RSA) and deep neural networks (DNNs) can be leveraged to gain insights into the structure of neural representations and their different representational formats. We provide several examples of recent studies which demonstrate that we are able to not only measure memory representations using RSA but are also able to investigate their multiple formats using DNNs. We demonstrate that in addition to slow generalization during consolidation, memory representations are subject to semantization already during short-term memory, by revealing a shift from visual to semantic format. In addition to perceptual and conceptual formats, we describe the impact of affective evaluations as an additional dimension of episodic memories. Overall, these studies illustrate how the analysis of neural representations may help us gain a deeper understanding of the nature of human memory.

Keywords Memory · Neural representations · Representational similarity analysis · Representational formats · Deep neural networks

Introduction: why should we assume representations?

When we think back to what we did yesterday, we are usually able to literally *picture* how a specific episode looked like, and perhaps also how it sounded, smelled, and felt. This ability to form a mental image or internal representation plays a crucial role for both re-experiencing the past and making plans for the future (Schacter and Addis 2007; Bonnici et al. 2012; Cheng et al. 2016; Brown et al. 2016). How is the sensory information about this episode transformed into a long-lasting neural memory trace? Will different aspects such as visual and abstract information be stored differently in memory? How can we measure the representational format of memories?

First of all: What is a representation? Described as early as 1904 by Richard Semon (e.g., Schacter 2001), most cognitive neuroscientists nowadays believe that mental representations of past and future episodes rely on a neural substrate that we can localize in the brain—on the "neural representation" of the represented episode (deCharms and Zador 2000; Shea 2018)—this notion has not always been accepted. Beginning with the "cognitive revolution" in the 1960s, cognitivism replaced behaviorism, a scientific movement trying to explain behavior not only without introspection, but also without assuming mental representations (Watson 1913; Skinner 1953; Egan 2014; Shea 2018; Newen and Vosgerau 2020). In contrast to behaviorism, cognitivists emphasized the importance of intentional states and mental

Rebekka Heinen and Anne Bierbrauer have contributed equally.

✉ Rebekka Heinen
rebekka.heinen@rub.de

[1] Department of Neuropsychology, Institute of Cognitive Neuroscience, Faculty of Psychology, Ruhr University Bochum, Universitätsstraße 150, 44801 Bochum, Germany

[2] Institute for Systems Neuroscience, Medical Center Hamburg-Eppendorf, Martinistraße 52, 20251 Hamburg, Germany

[3] Department of Cognitive Psychology, Institute of Cognitive Neuroscience, Faculty of Psychology, Ruhr University Bochum, Universitätsstraße 150, 44801 Bochum, Germany

contents for understanding cognitive functioning. According to this representational view, a mental representation consists of (1) a vehicle, i.e., a physical entity such as a population of neurons that is able to represent information, and (2) a content, i.e., the information about the outside world or about internal states that is carried by the vehicle (Fodor 2008; Roskies 2021). In addition—and critical for our review—this content can have (3) different representational formats: A given experience can be either represented conceptually or non-conceptually (Boghossian 1995). While conceptual representational formats are composed of semantic thoughts, non-conceptual formats rely on sensory aspects of an experience. Arguably, most “real-life” representations consist of both representational formats. For example, the representation of a visit to the ocean (content) comprises the fact that one was at a certain beach at a certain time (conceptual representational formats) and the feeling of sand beneath one's feet, the color of the water, and the heat of the sun (non-conceptual representational formats). The brain states carrying both types of information constitute the vehicles of mental representations. In this review, we will focus on these two types of formats—perceptual and conceptual.

The representational theory of the mind assumes that cognitive functioning consists of the formation and the transformation of mental representations. It will thus be important to develop methods to measure these representations and assessing their vehicle in the brain has become a core aim of contemporary cognitive neuroscience.

A case for internal representations

“A neural representation is a pattern of neural activity that stands for some environmental feature in the internal workings of the brain” (Vilarroya 2017, p. 4) and focuses on particular features in the world—i.e., neural representations have a representational content and involve a particular representational format (deCharms and Zador 2000). At early steps of sensory processing, neural representations involve representational formats that are more strongly correlated with external input than at later processing stages. For example, Hubel and Wiesel (1959) studied how the early visual cortex responds to bars at different angular directions. The striate cortex and other cortices at the beginning of the sensory processing hierarchy exhibit pronounced topographic organization, such that the patterns of activity are isomorphic with the external world (Poldrack 2021). At later processing steps, neural representations are less strongly driven by sensory inputs and more strongly shaped by cognitive operations. A famous example of such a representation occurs in an experiment that Tolman described in his book “Cognitive maps in rats and men” (1948): A rodent explores a maze and may find rewards when choosing the correct path. After some time, the reward path is blocked, and the rodent is offered several different alternative paths. Tolman could demonstrate that rodents took the shortest alternative path. This is indicative of an internal representation—in this case of relative spatial locations—that is referred to as “cognitive map”, as the behavior of the rodent cannot be solely explained by stimulus–response learning based on stimulus-outcome associations.

How can we measure and analyze neural representations?

Out of many ideas and possibilities how stimulus information is represented in neural structures, three prominent theories evolved which differ regarding the neural features containing representations. On the level of single neurons, the ‘rate coding’ hypothesis claims that the mean firing rate of each neuron carries information about stimuli (Adrian 1928; DeCharms and Zador 2000). The ‘temporal coding’ hypothesis posits that in addition to the mean firing rate the precise timing of spikes is crucial (DeCharms and Zador 2000; Gerstner and Kistler 2002; Gollisch and Meister 2008). We consider these coding schemes on the single unit level as “sparse” since they focus on coding by one or a few neurons (Axmacher et al. 2008; Reddy and Kanwisher 2006). In addition, the activity of large populations of neurons also carries information (Deadwyler and Hampson 1997; DeCharms and Zador 2000; Georgopoulos et al. 1986; Hebb 1949). This scheme of ‘population coding’ would be consistent with a large number of broadly tuned neurons that code for a given stimulus (Reddy and Kanwisher 2006).

At the population level, neural representations can be measured by decoding approaches which can be applied to various kinds of non-invasive data in human participants (most importantly, functional magnetic resonance imaging, fMRI, or electroencephalography, EEG). In contrast to univariate analysis techniques which reflect overall activity changes a commonly used way to assess neural representations in cognitive neuroscience is multivariate pattern analysis (MVPA). With the advent of MVPA it has become possible to extract representational contents and formats from distributed patterns of neural activity, e.g., voxel activity values in fMRI data or power values at various frequency bands, time points, and channels in EEG data (Naselaris et al. 2011; Hebart and Baker 2018; Kunz et al. 2018; Roskies 2021).

When two stimuli elicit similar overall activity levels and their informational content is reflected by the pattern of voxel activations instead, it may be impossible to find univariate activation differences. Therefore, MVPA aims at decoding the information that the patterns of activity carry about external stimuli (Haynes and Rees 2006; Kriegeskorte et al. 2008a; Mur et al. 2009; Haxby et al. 2014; Kragel et al. 2018). Even when brain regions are relevant for processing a large number of different stimuli, it thus becomes possible to differentiate neural representations of two stimuli based on their activation pattern (Mur et al. 2009; Raizada et al. 2010), which may reflect a neural population code (Kamitani and Tong 2005; Watrous et al. 2015; Kriegeskorte and Diedrichsen 2019).

The underlying assumption of MVPA is that neural representations can be characterized via high-dimensional state-spaces whose dimensions correspond to stimulus attributes, and that each individual representation corresponds to one point in this space (Haxby et al. 2014). The two most commonly used MVPA methods are pattern classification (Pereira et al. 2009) and representational similarity analysis (RSA; Kriegeskorte et al. 2008a; Kriegeskorte and Diedrichsen 2019).

RSA allows researchers to characterize the geometry of a representational space that can be based on various stimulus features (Kriegeskorte and Kievit 2013; Haxby et al. 2014; Kriegeskorte and Wei 2021; Roskies 2021). Importantly, RSA abstracts from the specific type of data that is investigated (e.g., fMRI or EEG) and consists of a matrix of similarities which quantifies the (dis)similarity between neural representations (Haxby et al. 2014). Hence, it becomes possible to analyze second-order similarities—i.e., the correspondence between two separate similarity matrices (RDMs)—of (1) neural representations measured in different brain regions, species, or modalities, (2) neural activity and behavioral outcomes, or (3) neural activity and computational models (Kriegeskorte et al. 2008a; Kriegeskorte and Kievit 2013; Haxby et al. 2014; Roskies 2021). In other words, RSA allows for an analysis of any kind of data pattern irrespective of the data format.

Implementing RSA requires the coding of neural activity as vectors, separately for the experimental conditions (e.g., for stimuli in an experiment). Afterwards, the representational distances between these vectors are calculated. The similarity or distance measures that are most often used are Pearson or Spearman correlations, or Euclidean or Mahalanobis distance. Higher similarity corresponds to lower representational distance and vice versa. The result is a representational dissimilarity matrix (RDM; Kriegeskorte et al. 2008a), i.e., a matrix that reflects the similarities or distances between every stimulus (or more generally, condition) with every other stimulus, resulting in a nxn matrix (e.g., stimulus x stimulus, condition x condition). Approaches like multidimensional scaling allow for a mapping of this high-dimensional representational space in lower-dimensional spaces (often 2D or 3D) in order to facilitate interpretation.

We now can extract information from the RDMs to characterize the underlying neural representations. First, self-similarity, sometimes also called representational fidelity or reliability (see Xue 2018 for review), refers to the similarity of brain patterns when the same stimulus is presented twice. Although self-similarity most commonly refers to the similarity of a stimulus compared to others (i.e., to non-self similarity), some studies use this term to denote the similarity between repetitions of the same stimulus (i.e., Xue et al. 2010). Here, we use this term to refer to the first case (within vs. between similarity). This tells us how faithful a neural representation reflects a given stimulus. Second, RDMs allow us to investigate the relationships between different stimuli, i.e., between-item similarity. This between-item similarity may reflect the features of a stimulus that are represented by a given brain region—i.e., two stimuli with similar low-level visual features, such as spatial frequencies or gratings, have similar representations in early visual cortices, while conceptual similarities lead to similar representations in association cortices (Kriegeskorte et al. 2008a, b). Based on these differences, RSA allows unraveling the representational format of neural representations. This method is highly flexible since many different features or conditions can be investigated in one experiment. One possible application is the investigation of human episodic memory, which we will describe next.

How do neural representations relate to memory?

An episodic memory can be conceived of as an internal representation of a previous experience (Goldman-Rakic 1995; Brewer et al. 1998; Cheng et al. 2016; Vilarroya 2017). At the neural level, it is widely assumed that memory representations are stored in memory traces or engrams—a term coined by Richard Semon in order to refer to learning-induced alterations of brain (micro-)structure (Semon 1904, 1909).

According to Semon, engrams are biological states that are objectively observable, which means that in principle, we can locate and manipulate them (Semon 1904, 1909; Josselyn et al. 2015; Kunz et al. 2018). Second, they represent specific memory contents and thus, when activated, lead to expression of this memory content (i.e., behaviorally measurable memory retrieval) (Liu et al. 2014a, b; Kunz et al. 2018). Moreover, they may be distributed within and across brain areas, an aspect that had been suggested by Lashley (1950) and was empirically supported for encoding by Haxby et al. (2001) and for retrieval by Brodt et al. (2018) more than 50 years later. Each engram corresponds specifically and uniquely to one particular memory, and this relationship is stable such that a given engram, when activated, should always elicit the same memory (Han et al. 2009; Liu et al. 2012, 2014a, b; Kunz et al. 2018). However, engrams or memory traces may also be transformed by various factors, such as time, memory consolidation, and novel learning (Dudai et al. 2015).

The formation of episodic memories requires the representation of the episode in a lasting memory trace (Xue 2018). In humans, various characteristics of memory representations have been associated with episodic memory performance: First, high amounts of self-similarity—i.e., of the memory representation of a particular content—predict subsequent memory (Xue et al. 2010; Visser et al. 2013). This result was found across various brain regions involving frontoparietal areas, the posterior cingulate cortex and sensory regions that are involved in processing the respective stimuli (Xue 2018). Self-similarity of memory representations may either refer to situations when a particular stimulus is encoded multiple times (encoding-encoding-similarity) or when encoding and retrieval of the same stimulus are compared (encoding-retrieval-similarity), and both measures predict memory accuracy (Xue et al. 2010; Xue 2018; Ten Oever et al. 2021). Thus, higher similarity between memory representations of the same stimulus seems to support (recognition) memory.

Interestingly, between-item similarity has been associated with memory performance as well, although not necessarily in a positive manner: Indeed, different theoretical frameworks and empirical results predict a memory advantage either for more distinct or for more similar memory representations of different items. Some studies found that stronger discrimination between different items (i.e., higher distinctiveness) supports memory (LaRocque et al. 2013; Xue 2018). The distinctiveness hypothesis is based on the idea that distinctiveness reduces possible interference with other, similar stimuli and thereby supports memory (Kılıç et al. 2017). The idea of distinctiveness is closely related to pattern separation in the hippocampus, a process by which similar memories are stored as distinct, non-overlapping representations (Bakker et al. 2008; Yassa and Stark 2011). An fMRI study confirmed that higher pattern distinctiveness in the hippocampus is indeed associated with better memory performance (LaRocque et al. 2013). In contrast, in perirhinal and parahippocampal cortex as well as in the amygdala, higher between-item similarity of neural representations benefits memory encoding, possibly because they are integrated into one unique episode that is distinct from other episodes (Visser et al. 2011, 2013; LaRocque et al. 2013; Bierbrauer et al. 2021). While these studies point to better memory performance with higher distinctiveness, there is also evidence that global pattern similarity—i.e., the similarity between different exemplars of the same concept—may support memory (Davis et al. 2014), even causing false alarms for new exemplars of the same concept (Wing et al. 2020).

These results support the idea that memory representations have multiple representational formats, whose representational 'geometries' (generalized or distinct) may exert different influences on memory encoding. In this review, we define visual/perceptual representational formats as reflecting visual stimulus features (e.g., their colors, textures, or shapes). Conversely, we define conceptual/semantic formats as reflecting semantic stimulus features including category information. How can we quantify the representational formats and measure the degree to which a stimulus might be represented in a visual or a conceptual format?

Using deep neural networks as models of representational formats

In recent years, the field of artificial intelligence has revolutionized our lives, with artificial neural network models (ANN) achieving near human-like performance in areas such as language translation (Popel et al. 2020) and car driving (Gupta et al. 2021) and even out-performing humans in various complex games such as chess (McGrath et al. 2022), Go (Silver et al. 2017), Starcraft (Vinyals et al. 2019) or Stratego (Perolat et al. 2022). If ANN models are able to perform on a human level, can we also utilize them to better understand our own brain processes?

To gain insight into the transformation of visual features into conceptual representations, convolutional Deep Neural Networks (cDNNs) from object recognition (Fig. 1A) have become models of choice. These models process image input through several convolutional layers, which are connected sparsely, up to fully-connected layers that assign a label to contents of the image. Strikingly, recent multivariate studies have found the same visual hierarchy and gradient in feature complexity in cDNNs trained on object recognition as observed in the brain (Leeds et al. 2013; Khaligh-Razavi and Kriegeskorte 2014; Güçlü and van Gerven 2015; Yamins and DiCarlo 2016; Cichy et al. 2016; Wen et al. 2018). These results were obtained across various data modalities, ranging from fMRI (Güçlü and van Gerven 2015; Allen et al. 2022) via magnetoencephalography (MEG; Clarke et al. 2018) and scalp EEG (Graumann et al. 2022) to oscillations in intracranial EEG (Kuzovkin et al. 2018) and monkey single-unit data (Cadieu et al 2014) and even behavioral outcomes, such as similarity judgements (Mur et al. 2013; Davis et al. 2021).

These findings demonstrate that the internal representations on multiple levels of complexity formed by cDNNs are closely linked to the features that are processed along the ventral visual stream (VVS) (Fig. 1A). Processing of visual information along the VVS reveals a hierarchy from basic visual features to higher-order visual and semantic category features (Cowell et al. 2010; DiCarlo et al. 2012; Kravitz et al. 2013). The visual cortex processes low-level visual properties, such as colors, shapes, and textures, and neurons in these areas have small receptive field sizes that

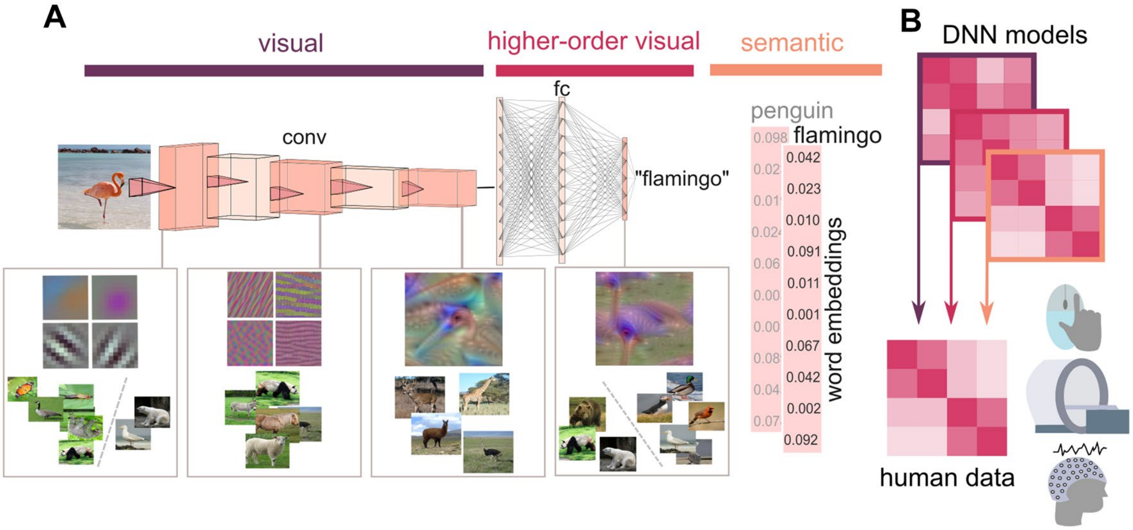

Fig. 1 Linking representational formats in DNNs and in semantic models to representations in the human brain. **A** Different representational formats in convolutional deep neural networks (cDNNs) and deep natural language processing models (dNLPs). In cDNNs, images are processed with a gradient in complexity, comparable to the human ventral visual pathway. As in the brain, basic visual information is processed in the first layer (whose receptive field properties roughly match those of V1) and is then passed through the convolutional layers of the network, which process increasingly complex information. After the last convolutional layer, the connections change, as now each neuron is connected to each neuron in subsequent layers (fully-connected). The network then chooses the most active (i.e., most likely) label in the highest layer. Early layers of the cDNN process edges, colors, and textures, such that e.g., animals of different categories (species) are sorted together e.g., based on their color. DNN neurons in middle layers have more complex receptive fields, processing object features such as the beak of the flamingo or their long necks and legs. Late layers respond to a visual prototype of the flamingo and show distinct representational similarity patterns (e.g., all bears are sorted together, even if they differ in their low-level visual features). While convolutional layers process lower-level sensory information, fully-connected layers process higher-order visual features including object classes. In addition to the cDNN, one can use dNLP models to quantify the representation of semantic information, based on embedding vectors of words or sentences. **B** The neural activations corresponding to all pairs of stimuli can be used to generate RDMs for all layers of the cDNN and the semantic dNLP, which can then be correlated to RDMs from brain or behavioral data generated using the same stimuli

lead to pronounced retinotopic specialization (Hubel and Wiesel 1962). As the signal progresses through the VVS to more anterior regions, such as the inferior temporal cortex (IT cortex; Kriegeskorte et al. 2008b), the fusiform gyrus (Clarke et al. 2011) and the lateral occipital cortex (LOC; Tyler et al. 2013), feature complexity and receptive field size increase, leading to higher-order representational formats involving object parts and domain-level semantic features (Clarke et al. 2013; Clarke 2015).

In cDNNs from object recognition, starting with low-level features such as edges and colors in the first convolutional layer, complexity increases to textures, object parts and finally, object categories in the last network layer. While early cDNN layers show similar activation patterns for images with shared visual features such as similar colors and textures (e.g., orange color of a pumpkin and of a basketball), independent of the conceptual similarity of stimuli, category-specific features explain similarities in later layers (e.g., wings, feathers and a beak for birds). Similar representational transformations have been found along the VVS (Mur et al. 2013; Hebart et al. 2020), revealing that cDNN models from computer vision can accurately reflect neural representations during object recognition. Surprisingly, in contrast to this functional overlap, the most prominently used network “AlexNet” (Krizhevsky et al. 2017) contains one of the simplest architectures. Yet, AlexNet and other, shallower cDNN models such as CorNet (Kubilius et al. 2018, 2019) are models that match neural representations relatively well (Nonaka et al. 2021) and show high classification performance in object recognition. On the other hand, it has been demonstrated that recurrency in DNN architectures may further improve the match to neural representations during object recognition (Kubilius et al. 2019; Kietzmann et al. 2019b; van Bergen and Kriegeskorte 2020), suggesting that recurrent DNNs should be increasingly used in the future to study neural representations.

Neural representations can be mapped onto DNN feature spaces via RSA, using the fact that RSA reflects representational geometry independent of data modality (Fig. 1B). Treating DNN feature activations as patterns allows one to compute similarities between all pairs of stimuli in each layer of the DNN or model, resulting in one RDM per DNN layer/model. Subsequently, DNN RDMs and neural RDMs can be correlated and compared for their similarity structures

(Mur et al. 2013; McClure and Kriegeskorte 2016). Several visualization techniques such as multi-dimensional scaling (MDS; Lin et al. 2019; Fig. 1A), class activation maps (Zhou et al. 2015) or similarities from RDMs (Kriegeskorte and Golan 2019) provide information on the representational formats that are processed in individual DNN layers or models. Linking DNN representations to neural representations thus allows one to examine properties of neural representations (e.g., in terms of brain regions, oscillation frequencies, or latencies) that reflect different representational formats and are for example responsible for the shift from perceptual to conceptual formats.

Many studies mentioned above employed cDNN models from image classification challenges to assess representational formats during neural processing. However, these cDNN models are limited to visual and higher-order visual representations, while category abstraction and many memory functions rely on conceptual representations (Clarke 2019). More specifically, even though perceptual features may allow for the derivation of conceptual representations in a feed-forward way (Clarke et al. 2018), this process can be facilitated by top-down semantic knowledge (Taylor et al. 2012; van Kesteren et al. 2013), emphasizing the need for models that involve semantic processing. In fact, research on language processing even provided evidence for multiple levels of semantic features, as indicated by faster performance for general domain features compared to exemplar-specific features (Randall et al. 2004; Macé et al. 2009; Devereux et al. 2018). Thus, instead of focusing on one single cDNN model, Clarke (2019) proposed the additional use of deep learning models from natural language processing (deep Natural Language Processing models; dNLP). Previous research showed that these models can accurately reflect conceptual representations during object recognition in the VVS (Devereux et al. 2018) and even during more abstract tasks such as those involving narrative content (Lee and Chen 2022). DNLP models are trained on text input (e.g., wiki pages, books, user reviews) rather than images. These corpus-based models, such as BERT (Devlin et al. 2018), the Google Sentence encoder (Cer et al. 2018), Infersent (Conneau et al. 2017) or GPT-3 (Brown et al. 2020) assign a word embedding vector to each word or sentence based on co-occurrences of these concepts, and these vectors can then be used to study semantic similarities.

Taken together, although cDNN models are only very rough approximations to the neural processes and connectivity within the VVS, the findings reviewed above demonstrate that representations in DNN layers are relatively accurate models of the neural representations at different levels of abstraction, which makes them specifically interesting to study neural properties of and changes in representational format (Marblestone et al. 2016; Kietzmann et al. 2019a; Richards et al. 2019; Storrs and Kriegeskorte 2019; Saxe et al. 2021). Surprisingly, only few studies thus far employed DNNs to study representational formats of memory representations. Davis et al. (2021) were among the first to apply visual and semantic DNN model features to investigate the effects of representational formats during encoding on subsequent memory. Participants first viewed images of natural objects that they had to name and were then tested in two retrieval tasks. During retrieval, the authors separately made either perceptual or conceptual formats task-relevant by either displaying old and new images (perceptual) or the label of the concepts (conceptual). Using fMRI, the authors could show that matching of encoding representations to RDMs from either a visual cDNN or semantic models (taxonomy/encyclopedic) predicted memory performance in both retrieval conditions. Conceptual and perceptual formats recruited different brain areas though, namely the anterior VVS and the early visual cortex, respectively. Interestingly, although the two representational formats were linked to different brain areas depending on the retrieval task (perceptual/conceptual), the performance in both tasks benefited from matching with the respective other format during retrieval as well, suggesting that perceptual memory benefits from top-down information, while bottom-up visual information facilitates conceptual memory.

The role of representational formats for understanding the dynamics of memory representations

Several findings of item-specific memory representations concern frontoparietal and midline regions (e.g., Baldassano et al. 2017; Fernandino et al. 2022; Huth et al. 2012; Lee and Kuhl 2016). Since these areas do not reflect sensory processing, it is currently not clear why they exhibit pronounced stimulus specificity. In the future, DNNs trained on more complex objective functions than stimulus categorization may account for the formats in these areas. However, studies using RSA identified an important role of the VVS in transforming visual into conceptual representations (DiCarlo et al. 2012; Kravitz et al. 2013; Martin et al. 2018). This transformation from perceptual to conceptual representations during perception (Kriegeskorte and Kievit 2013) may also give rise to different representational formats of memory traces, which may rely predominantly on either perceptual or semantic representational formats as well. Since both visual and semantic formats play a role during object recognition, the question arises whether memory traces during the different stages of memory processing—encoding, short-term memory maintenance, consolidation, and retrieval—would reveal such formats as well. Already at early visual processing steps, top-down knowledge plays an important role. Typically, we do not encounter objects without any prior information on their use and behavioral importance but using conceptual representations that are stored in long-term memory (Tulving and Watkins 1975; Xue 2018). At the same time, neural representations are not stable but subject to transformation processes (Xue 2022). According to the neural-psychological-representation-correspondence (NPRC) by Gilboa and Moscovitch (2021) memory traces can occur in different forms, a given episode can be represented in an event-specific visual format, while at the same time containing information about schemas and semantic information from prior knowledge. In addition, these representational formats may dynamically change due to various factors such as time after encoding, task context, goals, or prior knowledge, resulting in transformations between formats. How are these representations formed and especially, how are they transformed?

One candidate framework on the transformation of visual information into long-term memory is based on the concept of semantization or gist-abstraction (Konkle et al. 2010; Winocur and Moscovitch 2011; Linde-Domingo et al. 2019; Lifanov et al. 2021). During semantization, sensory information is integrated into long-term semantic knowledge through representational transformations (Paller and Wagner 2002; Xue 2018; Favila et al. 2020). According to this framework, conceptual features of a sensory input are selectively strengthened, while detailed sensory information is reduced, facilitating the integration of novel experiences with prior semantic knowledge. In line with this theory, studies found better memory performance for conceptual features as compared to low-level/perceptual features (Bainbridge et al. 2017; Bainbridge 2019; Linde-Domingo et al. 2019). Memory was also improved for stimuli that could easily be linked with pre-existing schemas as compared to those that did not match a schema (van Kesteren et al. 2013), and reaction times were faster for conceptual compared to perceptual features during recall (Lifanov et al. 2021). Thus, semantization can be defined as a transformation from detail-rich to compressed gist-like representations, suggesting a change in representational format. In addition, one would expect a transformation of memory traces such that they become more similar for stimuli that share the same prototypical conceptual features (e.g., beak and wings of birds) and less similar for stimuli with similar visual details (e.g., red parrot and red tomato).

In this case, semantization may actually lead to an increase of false alarms to semantically similar lures or to novel exemplars of previously presented concepts. Indeed, Naspi et al. (2021a) found more errors for lures consisting of prototypical exemplars of a given category, indicating that enhanced gist-abstraction during encoding or consolidation can lead to increased false alarms at recognition. At the same time, false alarms also increased for lures with high visual similarity to originally encoded images, suggesting that not all unique visual information is lost after encoding. Delhaye and Bastin (2021), who focused on the impact of visual or semantic processing during encoding, found semantization to be independent of encoding type format. Interestingly, Naspi et al. (2021b) could show that both visual and semantic formats in VVS contributed to successful memory encoding, but categorical information in regions anterior to the VVS predicted later forgetting. These studies demonstrate that there is no rigid transformation from visual to semantic formats during encoding but an interplay between different formats at different steps of memory processing.

A substantial body of evidence has shown off-line replay of memory representations during sleep (Frankland and Bontempi 2005; Deuker et al. 2013; Dudai et al. 2015). Integration of novel experiences into prior knowledge is assumed to be caused by strengthening of those features that are shared across encoded contents (Káli and Dayan 2004; Lewis and Durrant 2011; Himmer et al. 2019). These results are in line with the idea that replay facilitates generalization processes (Liu et al. 2019). While sleep contributes to integration by enhancing memory for shared features of newly encoded content, there is also evidence for sleep to prevent loss of unique feature representations (Schapiro et al. 2017). This might indicate that not only conceptual but multiple representational formats, including perceptual details, are strengthened due to off-line replay during sleep, which in turn might slow down the supposed loss of visual detail over time.

Perceptual details might be subject to faster forgetting in order to promote an integration of conceptual or superordinate categorical features into memory, considering that different representational formats of a memory trace may be forgotten independently (Brady et al. 2013). In line with supposed gist abstraction and loss of visual detail due to semantization, Lifanov et al. (2021) found that a perceptual-conceptual gap (e.g., a shift from faster reaction times for perceptual features to faster reaction times for conceptual features) increased over time, suggesting faster forgetting of visual details while conceptual features were integrated into long-term memory (LTM). During retrieval, conceptual features were activated prior to visual detail when no visual input was present (Linde-Domingo et al. 2019; Davis et al. 2021) and were found to be involved during memory retrieval of both perceptual and conceptual representational formats (Davis et al. 2021; see above). While these results could lead to the wrong conclusion of a complete loss of visual detail over time, Ferreira et al. (2019) found higher neural similarities between category-related but also between episode-unique information, demonstrating that conceptualization during semantization does not necessarily come at the cost of visual detail.

Overall, these findings deliver further evidence for a dynamic transformation of representational formats (Paller and Wagner 2002; Xue 2018, 2022; Favila et al. 2020; Liu et al. 2020, 2021). Yet, the question remains whether consolidation induces transformations of memory traces from one format to another (i.e., from perceptual to conceptual, losing all perceptual detail) or whether memory traces consist of multiple formats with only their accessibility changing across time, and depending on encoding tasks and/or retrieval cues. While this can be investigated by analyses of encoding-retrieval similarity (i.e., Ten Oever et al. 2021), DNNs could be used to further investigate the underlying representational formats and to address the question whether these formats are subject to transformation or continue to coexist. In the next section of this review, we will take a closer look at how DNNs may be used to investigate such changes in the representational format of memory traces during the earliest possible stage when semantization might occur, i.e., directly after the offset of a stimulus, and during subsequent processing stages.

Beyond recognition—using DNNs to investigate early stages of semantization

As described above, previous research has demonstrated that cDNN features can be used to study representational formats during object recognition (Güçlü and van Gerven 2015; Cichy et al. 2016; Kuzovkin et al. 2018; Clarke et al. 2018). However, these studies did not assess the question of how these visual inputs were transformed during the consecutive stages of long-term memory encoding, memory consolidation, and long-term memory retrieval. Investigating representational formats during initial stages of memory formation, we could test whether the supposedly slow process of gist-abstraction (O'Reilly et al. 2014) unfolding during systems consolidation might happen more rapidly and already in earlier post-encoding stages.

A very recent study thus set out to investigate if DNN similarities would reflect neural similarities during a visual short-term memory (VSTM) task (Fig. 2A), with VSTM being the earliest offline processing stage following perception (Liu et al. 2020). A follow-up study (Fig. 2B) then tested the effects of short-term maintenance and consecutive transformation stages on LTM retrieval (Liu et al.

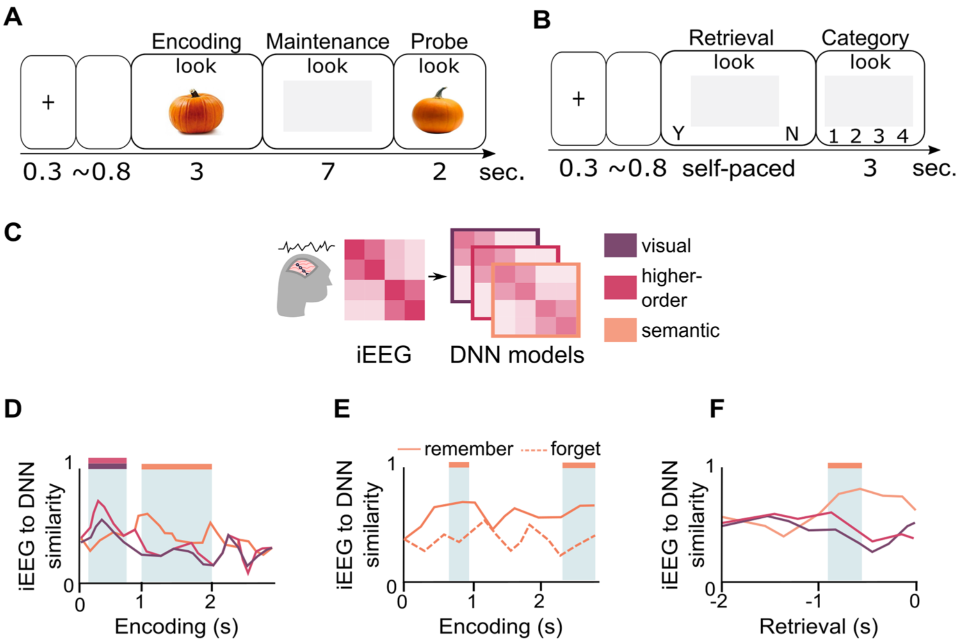

Fig. 2 Representational formats during visual short-term memory maintenance and long-term memory encoding and retrieval. **A** Participants saw cue-object pairs and maintained the objects during a long maintenance period, which was followed either by a picture of the same item or of a similar lure. **B** During a subsequent long-term memory test, the cue word was presented, and participants were asked to vividly imagine the associated image. Afterwards they conducted a forced choice test on the category of the image. **C** Analysis methods: We combined RDMs from eight different layers of a cDNN and from a dNLP to model lower-order visual, higher-order visual and semantic representational formats, respectively. Model RDMs were then compared to corresponding RDMs from intracranial EEG data during the different task periods. **D** During encoding, higher-order visual formats were gradually transformed into semantic representational formats. **E** More pronounced semantic formats during encoding predicted subsequent long-term memory success. **F** Semantic but not visual formats were found during successful memory retrieval (s=seconds)

2021) to test whether semantization already occurs during VSTM and whether early semantization may improve LTM performance.

VSTM is defined as the active maintenance of visual information for a short period of time in a limited capacity store (Baddeley and Hitch 1974; Luck and Vogel 1997). Current research suggests an important role of VSTM for integrating information, bridging the gap between perception and long-term memory (Chota and Van der Stigchel 2021), specifically involving regions along the VVS (Meyers et al. 2008; Cichy et al. 2014). Studies indicate "dynamic coding" with neurons carrying different information across the maintenance period (Stokes 2015), reflected by distinct representational formats. Along the VVS, these distinct formats have already been observed during object recognition (e.g., Devereux et al. 2018). Is there evidence for different representational formats present already during VSTM and for a shift from perceptual to semantic formats prior to LTM retrieval?

To address this question, we first analyzed similarities of neural patterns during an encoding and a maintenance period in a delayed matching to sample task (Fig. 2A) while participants (presurgical epilepsy patients) underwent intracranial EEG (iEEG) recordings (Liu et al. 2020). We then examined whether neural patterns during encoding reappeared during maintenance and long-term memory retrieval (Liu et al. 2021). During the maintenance period of the VSTM task, we found item-specific reinstatement of information from two distinct time windows during encoding, an early (250–770 ms post stimulus onset) and a later period (1000–1980 ms post stimulus onset), suggesting that both periods may contain different representational formats. Further analyses revealed higher item-specificity for the late encoding time window, indicating that specifically late representational formats are maintained faithfully during VSTM. Thus, neural similarity analysis revealed reinstatement of two distinct formats, but how exactly can these formats be characterized?

The integration of visual input with long-term knowledge suggests an involvement of semantic information (Cichy et al. 2014; Stokes 2015), while cDNNs from object recognition capture visual and higher-order visual features only. Thus, we decided to combine a cDNN with a dNLP model to investigate matching of neural representational formats to either visual formats from the cDNN or semantic formats from the dNLP model (Fig. 2C).

Current theories suggest an involvement of VSTM in the transformation of visual stimuli into abstract long-term memory representations (Meyers et al. 2008; Cichy et al. 2014; Stokes 2015). Accordingly, we found evidence for visual features during stimulus encoding periods followed by abstract semantic representations during later processing periods, indicating a transformation of representational formats from sensory to abstract (i.e., non-perceptual) formats (Fig. 2D). Specifically, the absence of sensory information during the later period suggests an interplay between bottom-up visual processing during the early and semantic top-down processing during later processing steps—possibly reflecting the integration of novel sensory stimuli into long-term memory stores (Clarke 2015; Jozwik et al. 2017; O'Donnell et al. 2018). In addition, the presence of a semantic format may be beneficial in order to transform stimuli into a lower-dimension representation with reduced information content, which may provide a functional benefit: Conci et al. (2021) found VSTM capacity to be linked to participants' prior knowledge, with higher capacity if the stimulus meaning was known.

Recent findings from studies using fMRI provide additional evidence of shared representational formats between VSTM and long-term memory retrieval (Vo et al. 2022). Bainbridge et al. (2021) found different levels of abstraction when comparing encoding and retrieval representations. Whereas both fine-grained (e.g., penguin, lion) and coarse (e.g., bird, feline) features were observed during encoding, primarily coarse features were present during recall. Specifically, their results demonstrate a shift from the VVS showing peak activity during encoding to anterior areas during retrieval. Yet, this study does not show a complete loss of perceptual (e.g., fine-grained) information that would reflect a transformation of the same memory trace since this perceptual information was still observable in some areas. Whereas Audrain and McAndrews (2022) also found that memory representations became coarser over time, interestingly they found this generalization was linked to prior knowledge with only congruent semantic stimuli associations integrated in the medial prefrontal cortex (mPFC). This suggests that rapid semantization, i.e., due to congruency to prior knowledge, can facilitate memory transformation.

A follow-up analysis on the results during VSTM described above supports these findings even further (Liu et al. 2021): In line with our results from the maintenance period, we found that transformation into semantic formats was linked to subsequent LTM performance. Specifically, remembered images showed more pronounced semantic formats during encoding compared to forgotten images (Fig. 2E) and were linked to the occurrence of semantic but not sensory formats during retrieval (Fig. 2F). Interestingly, item-specific memory representations during retrieval were more similar to the visual short-term maintenance period compared to encoding. Together with better memory performance when conceptual formats were abstracted during encoding, the findings of these studies suggest semantization already happening at early stages of memory (i.e., encoding and VSTM) which in turn leads to better long-term memory formation. Overall, there seem to be parallel generalization processes during both encoding and consolidation, modulated by factors such as prior knowledge, which are fundamental to memory formation in both humans and neural networks (Kumaran et al. 2016) and supposedly are not limited to post-encoding consolidation periods.

Beyond sensory formats—affective and contextual dimensions

In previous sections we focused on perceptual and conceptual formats, yet we hardly believe that these two cover the entirety of representational dimensions in neurocognitive processing (Gilboa and Moscovitch 2021). There might be additional, more abstract formats (e.g., involving scripts and schemata) or additional dimensions, such as the affective evaluation or the contextual embedding of an episode. When we think back to the example described above on a day at the beach, we may not only remember its multisensory aspects (e.g., the feeling of the sand, the sound of waves) but also the emotions we felt in that moment.

Indeed, there is evidence for neural representations of affective dimensions and categories across large-scale brain networks (Kragel and LaBar 2016), even spanning to areas along the VVS (Kragel et al. 2019). Concerning memory representations of emotional contents and their potential contextual embedding, emotions have been shown to modulate memory formation via processes of emotional binding (Mather 2007; Yonelinas and Ritchey 2015). Typically, emotions enhance memory (Talmi 2013; LaBar and Cabeza 2006), and this is particularly the case for negative emotions (Kensinger 2007). Interestingly, negative emotions seem to specifically enhance certain representational formats, with some studies indicating better accessibility of perceptual formats (Kensinger et al. 2006, 2007). Similar effects may occur for negative emotions induced by psychosocial stress, a particularly ecologically relevant condition (Freund et al. 2023). How will affective evaluation and contextual embedding affect neural representations of different stimuli of a stressful episode?

It is well established that the effects of stress on memory depend on the phase (Roozendaal 2002; Het et al. 2005; Joëls et al. 2006; Wolf 2009; Shields et al. 2017) of memory processing. While stress before or during encoding may have mixed effects, it is usually beneficial when experienced after encoding or during consolidation. By contrast, experiencing stress shortly before or during retrieval is consistently detrimental to performance (de Quervain et al. 1998; Wolf 2009; Shields et al. 2017).

In order to investigate memories of a stressful episode, we applied an ecologically valid experimental design in which stimuli are incidentally encoded during a psychosocial stress intervention (Wolf 2019). In the Trier Social Stress Test (TSST; Kirschbaum et al. 1993), participants conduct a mock job interview in front of a neutrally acting committee. In previous studies, Wiemers et al. adapted the TSST to contain a number of different everyday objects (Wiemers et al. 2013; Wolf 2019). In this version of the TSST (Fig. 3A), the interview room and especially the table in front of the committee are equipped with a number of different objects, which are incidentally encoded during the TSST. Half of these objects are manipulated by the committee members in a standardized way to render them more salient for the participant ("central objects"), while other objects are not manipulated ("peripheral objects"). Several studies showed that central objects are better remembered than peripheral objects, and that this effect is increased by psychosocial stress (Wiemers et al. 2013, 2014; Herten et al. 2017a, b).

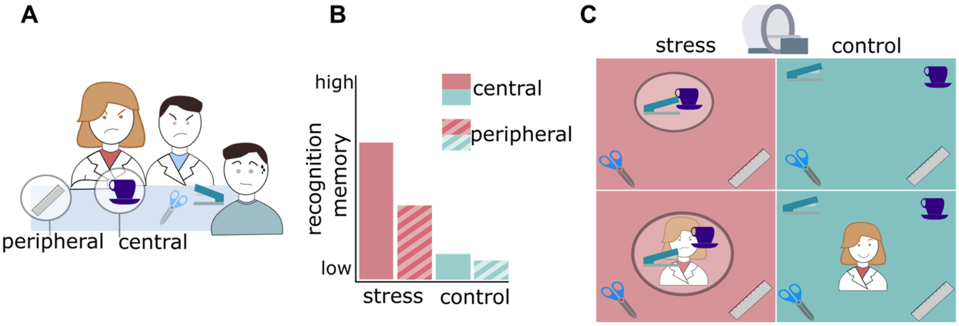

Fig. 3 Effects of psychosocial stress on memory representations. **A** Participants conducted a psychosocial stress intervention (Trier Social Stress Test, TSST) in which some objects were manipulated by stress-inducing committee members (central objects) while others were not (peripheral objects). A second group of participants took part in a non-stressful control version of the task. **B** We found that central objects were better remembered than peripheral objects, and that this effect was enhanced in stressed participants. **C** On the next day, pictures of all objects were presented in the MRI scanner, and we measured their representational similarity. In the left amygdala, we found that central objects were more similar to other objects of the same episode and dissimilar to distractor objects when comparing stressed vs. control participants (upper row). Better memory performance for these objects was explained by the similarity of their representations to the representation of the stressor, i.e., the committee members' faces (lower row)

These results are in line with the hypothesis that stress particularly enhances the encoding of central cues (Easterbrook 1959; Wolf 2019).

We speculated that stress may enhance later memory for central objects by supporting generalization or binding processes of their neural representations. The term "binding" typically refers to the formation of integrated representations of multiple aspects of an episode, i.e., of different elements within one spatiotemporal context, and has been proposed to rely critically on the hippocampus (Ranganath 2010; Eichenbaum 2017). Yonelinas and Ritchey (2015) have suggested an "emotional binding" account according to which an emotion instead of the spatio-temporal context binds the features of an episode. They proposed that emotional binding occurs in the amygdala, that it may outweigh spatio-temporal binding processes in the hippocampus, and that this is the reason why emotional memories are less likely to be forgotten. Another binding approach proposed by Mather (2007) may provide an explanation for the superior memory of the central aspects in an emotional episode. In her "object-based framework", she suggests that emotionally arousing objects attract attention and that this is the reason why the constituent features of the object are bound and well-remembered.

On the neural level, these binding approaches would predict higher similarity (lower representational distances) between neural representations of central objects. Generalization effects in humans have been previously found in a fear learning paradigm and were predictive of long-term fear memories (Visser et al. 2011, 2013). Specifically, pattern similarity changes in ventromedial PFC at the time of learning could predict the behavioral expression of long-term fear learning, i.e., changes in pupil dilation (Visser et al. 2013). In addition, fear learning led to generalization in other brain regions such as anterior cingulate cortex, amygdala, and superior frontal gyrus (Visser et al. 2011). These results suggest that increased pattern similarity between conditioned and unconditioned stimuli supports fear conditioning in a variety of brain regions including the amygdala—i.e., that higher pattern similarity in these regions reflects generalization and binding processes.

We investigated the effects of stress on memory representations and their impact on subsequent recognition memory (Bierbrauer et al. 2021). We conducted the TSST (Fig. 3A) and a non-stressful control version and tested memory performance for central and peripheral objects and the faces of the committee members. In line with previous studies (Wiemers et al. 2013), central objects were generally better remembered than peripheral objects (Fig. 3B). This effect was significantly more pronounced for stressed participants. Using fMRI, we measured the neural representations of central and peripheral objects and of the faces (Fig. 3C). Interestingly, we found that neural representations of central objects in the stressful episode became more similar to other objects from the same episode and dissimilar to distractor objects (i.e., objects that belonged to other potential episodes). In addition, we could explain higher memory performance for these objects by the similarity of their representations to the representation of the stressor, i.e., the committee members' faces. This suggests that the beneficial effects of stress on memory formation rely on a generalization of neural representations within the stressful episode, which is driven by higher similarity with the representation of the stressor. This representational change may also explain why memories of stressful experiences can be triggered by neutral cues with low representational distance to the stressor.

Our study demonstrates that investigating the representational structure or "geometry" of affective and contextual dimensions of memory traces may provide mechanistic insights into representational formats beyond perceptual and conceptual dimensions. In other words, understanding how neural representations are transformed by factors such as stress will help us understand how these factors change our memories. In the future, it would be interesting to link perceptual and conceptual format from DNNs to data from affective episodes to further broaden the understanding of how affective evaluation and contextual embedding modulates these formats, and how they may act as additional formats.

Conclusions

We started out describing several aspects of the memory for a recently experienced episode. The mental "image" of this episode as well as its non-sensory aspects relate to neural representations in various brain regions, across several levels of brain organization and in different representational formats. We described that RSA is well suited to assess the structure of memory representations (i.e., their representational geometry) and that we can employ DNNs to differentiate multiple representational formats. Not only can we quantify the formats themselves, but we also gain insights into how one format is transformed into another format and how this process may benefit memory consolidation and long-term memory retrieval. Importantly, we can demonstrate that generalization is not limited to consolidation but may also happen more rapidly, i.e., during encoding and maintenance. In this review we highlighted visual, higher-order visual and semantic formats that can be easily modeled by current cDNN and dNLP architectures. These models only provide a first approximation to the large variety of representational formats that are processed in the brain, including formats along dimensions such as affective evaluation or contextual embedding. In addition, they indicate the importance of combining computational and neuroscientific methods to understand memory. We propose that elucidating the neural representations underlying episodic memories should be a major goal in memory research.

Author contributions RH: Conceptualization, writing—original draft, writing—review and editing, visualization, project administration. AB: Conceptualization, writing—original draft, writing—review and editing. OTW: Resources, writing—review and editing. NA: Conceptualization, resources, writing—review and editing, supervision, project administration.

Funding Open Access funding enabled and organized by Projekt DEAL. R.H. and A.B. received funding from the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation)—project no. 122679504—SFB 874. O.T.W. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project no. 122679504—SFB 874, project no. 419039274—FOR 2812 and project no. 396877307 (ORA framework). N.A. received funding from the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation)—project Nos. 316803389—SFB 1280, via project no. 122679504—SFB 874, and via ERC CoG 864164.

Data availability Enquiries about data availability of the experiments presented in this review should be directed to the corresponding author.

Declarations

Conflict of interest The authors declare that they have no competing interests.

References

Adrian EDA (1928) The Basis of Sensation, the Action of the Sense Organs. Norton, New York

Allen EJ, St-Yves G, Wu Y et al (2022) A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nat Neurosci 25:116–126. https://doi.org/10.1038/s41593-021-00962-x

Audrain S, McAndrews MP (2022) Schemas provide a scaffold for neocortical integration of new memories over time. Nat Commun 13:5795. https://doi.org/10.1038/s41467-022-33517-0

Axmacher N, Elger CE, Fell J (2008) Memory formation by refinement of neural representations: the inhibition hypothesis. Behav Brain Res 189:1–8. https://doi.org/10.1016/j.bbr.2007.12.018

Baddeley AD, Hitch G (1974) Working memory. Psychology of learning and motivation. Elsevier, pp 47–89

Bainbridge WA (2019) Memorability: how what we see influences what we remember. Psychology of learning and motivation. Academic Press, pp 1–27

Bainbridge WA, Dilks DD, Oliva A (2017) Memorability: a stimulus-driven perceptual neural signature distinctive from memory. Neuroimage 149:141–152. https://doi.org/10.1016/j.neuroimage.2017.01.063

Bainbridge WA, Hall EH, Baker CI (2021) Distinct representational structure and localization for visual encoding and recall during visual imagery. Cereb Cortex 31:1898–1913. https://doi.org/10.1093/cercor/bhaa329

Bakker A, Kirwan CB, Miller M, Stark CEL (2008) Pattern separation in the human hippocampal CA3 and dentate gyrus. Science 319:1640–1642. https://doi.org/10.1126/science.1152882

Baldassano C, Chen J, Zadbood A et al (2017) Discovering event structure in continuous narrative perception and memory. Neuron 95:709-721.e5. https://doi.org/10.1016/j.neuron.2017.06.041

Bierbrauer A, Fellner M-C, Heinen R et al (2021) The memory trace of a stressful episode. Curr Biol 31:5204-5213.e8. https://doi.org/10.1016/j.cub.2021.09.044

Boghossian P (1995) Content. In: Kim J, Sosa E, R RS (eds) Companion to metaphysics. Oxford, pp 94–96

Bonnici HM, Chadwick MJ, Lutti A et al (2012) Detecting representations of recent and remote autobiographical memories in vmPFC and hippocampus. J Neurosci 32:16982–16991. https://doi.org/10.1523/JNEUROSCI.2475-12.2012

Brady TF, Konkle T, Alvarez GA, Oliva A (2013) Real-world objects are not represented as bound units: independent forgetting of different object details from visual memory. J Exp Psychol Gen 142:791–808. https://doi.org/10.1037/a0029649

Brewer JB, Zhao Z, Desmond JE et al (1998) Making memories: brain activity that predicts how well visual experience will be remembered. Science 281:1185–1187. https://doi.org/10.1126/science.281.5380.1185

Brodt S, Gais S, Beck J, Erb M, Scheffler K, Schönauer M (2018) Fast track to the neocortex: a memory engram in the posterior parietal cortex. Science 362(6418):1045–1048. https://doi.org/10.1126/science.aau2528

Brown TI, Carr VA, LaRocque KF et al (2016) Prospective representation of navigational goals in the human hippocampus. Science 352:1323–1326. https://doi.org/10.1126/science.aaf0784

Brown TB, Mann B, Ryder N et al (2020) Language models are few-shot learners. Adv Neurol Inform Process Syst. arXiv:2005.14165

Cadieu CF, Hong H, Yamins DLK et al (2014) Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol 10:e1003963. https://doi.org/10.1371/journal.pcbi.1003963

Cer D, Yang Y, Kong S-Y et al (2018) Universal sentence encoder. arXiv:1803.11175

Cheng S, Werning M, Suddendorf T (2016) Dissociating memory traces and scenario construction in mental time travel. Neurosci Biobehav Rev 60:82–89. https://doi.org/10.1016/j.neubiorev.2015.11.011

Chota S, Van der Stigchel S (2021) Dynamic and flexible transformation and reallocation of visual working memory representations. Vis Cogn 29:409–415. https://doi.org/10.1080/13506285.2021.1891168

Cichy RM, Pantazis D, Oliva A (2014) Resolving human object recognition in space and time. Nat Neurosci 17:455–462. https://doi.org/10.1038/nn.3635

Cichy RM, Khosla A, Pantazis D et al (2016) Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep 6:27755. https://doi.org/10.1038/srep27755

Clarke A (2015) Dynamic information processing states revealed through neurocognitive models of object semantics. Lang Cogn Neurosci 30:409–419. https://doi.org/10.1080/23273798.2014.970652

Clarke A (2019) Neural dynamics of visual and semantic object processing. Psychol Learn Motiv 70:71–95. https://doi.org/10.1016/bs.plm.2019.03.002

Clarke A, Taylor KI, Tyler LK (2011) The evolution of meaning: spatio-temporal dynamics of visual object recognition. J Cogn Neurosci 23:1887–1899. https://doi.org/10.1162/jocn.2010.21544

Clarke A, Taylor KI, Devereux B et al (2013) From perception to conception: how meaningful objects are processed over time. Cereb Cortex 23:187–197. https://doi.org/10.1093/cercor/bhs002

Clarke A, Devereux BJ, Tyler LK (2018) Oscillatory dynamics of perceptual to conceptual transformations in the ventral visual pathway. J Cogn Neurosci 30:1590–1605. https://doi.org/10.1162/jocn_a_01325

Conci M, Kreyenmeier P, Kröll L et al (2021) The nationality benefit: long-term memory associations enhance visual working memory for color-shape conjunctions. Psychon Bull Rev 28:1982–1990. https://doi.org/10.3758/s13423-021-01957-2

Conneau A, Kiela D, Schwenk H et al (2017) Supervised learning of universal sentence representations from natural language inference data. arXiv:1705.02364

Cowell RA, Bussey TJ, Saksida LM (2010) Components of recognition memory: dissociable cognitive processes or just differences in representational complexity? Hippocampus 20:1245–1262. https://doi.org/10.1002/hipo.20865

Davis T, Xue G, Love BC et al (2014) Global neural pattern similarity as a common basis for categorization and recognition memory. J Neurosci 34:7472–7484

Davis SW, Geib BR, Wing EA et al (2021) Visual and semantic representations predict subsequent memory in perceptual and conceptual memory tests. Cereb Cortex 31:974–992. https://doi.org/10.1093/cercor/bhaa269

de Quervain DJ, Roozendaal B, McGaugh JL (1998) Stress and glucocorticoids impair retrieval of long-term spatial memory. Nature 394:787–790. https://doi.org/10.1038/29542

Deadwyler SA, Hampson RE (1997) The significance of neural ensemble codes during behavior and cognition. Annu Rev Neurosci 20:217–244

deCharms RC, Zador A (2000) Neural representation and the cortical code. Annu Rev Neurosci 23:613–647. https://doi.org/10.1146/annurev.neuro.23.1.613

Delhaye E, Bastin C (2021) Semantic and perceptual encoding lead to decreased fine mnemonic discrimination following multiple presentations. Memory 29:141–145. https://doi.org/10.1080/09658211.2020.1849309

Deuker L, Olligs J, Fell J et al (2013) Memory consolidation by replay of stimulus-specific neural activity. J Neurosci 33:19373–19383. https://doi.org/10.1523/JNEUROSCI.0414-13.2013

Devereux BJ, Clarke A, Tyler LK (2018) Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. Sci Rep 8:10636. https://doi.org/10.1038/s41598-018-28865-1

Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805

DiCarlo JJ, Zoccolan D, Rust NC (2012) How does the brain solve visual object recognition? Neuron 73:415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Dudai Y, Karni A, Born J (2015) The consolidation and transformation of memory. Neuron 88:20–32. https://doi.org/10.1016/j.neuron.2015.09.004

Easterbrook JA (1959) The effect of emotion on cue utilization and the organization of behavior. Psychol Rev 66:183–201. https://doi.org/10.1037/h0047707

Egan F (2014) How to think about mental content. Philos Stud 170:115–135. https://doi.org/10.1007/s11098-013-0172-0

Eichenbaum H (2017) On the integration of space, time, and memory. Neuron 95:1007–1018. https://doi.org/10.1016/j.neuron.2017.06.036

Favila SE, Lee H, Kuhl BA (2020) Transforming the concept of memory reactivation. Trends Neurosci 43:939–950. https://doi.org/10.1016/j.tins.2020.09.006

Fernandino L, Tong J-Q, Conant LL et al (2022) Decoding the information structure underlying the neural representation of concepts. Proc Natl Acad Sci USA. https://doi.org/10.1073/pnas.2108091119

Ferreira CS, Charest I, Wimber M (2019) Retrieval aids the creation of a generalised memory trace and strengthens episode-unique information. Neuroimage 201:115996. https://doi.org/10.1016/j.neuroimage.2019.07.009

Fodor JA (2008) LOT2: the language of thought revisited. Oxford University Press, Cambridge

Frankland PW, Bontempi B (2005) The organization of recent and remote memories. Nat Rev Neurosci 6:119–130. https://doi.org/10.1038/nrn1607

Freund IM, Peters J, Kindt M, Visser RM (2023) Emotional memory in the lab: using the Trier Social Stress Test to induce a sensory-rich and personally meaningful episodic experience. Psychoneuroendocrinology 148:105971

Georgopoulos AP, Schwartz AB, Kettner RE (1986) Neuronal population coding of movement direction. Science 233(4771):1416–1419. https://doi.org/10.1126/science.3749885

Gerstner W, Kistler WM (2002) Spiking neuron models: single neurons, populations, plasticity. Cambridge University Press, Cambridge, UK

Gilboa A, Moscovitch M (2021) No consolidation without representation: correspondence between neural and psychological representations in recent and remote memory. Neuron 109:2239–2255. https://doi.org/10.1016/j.neuron.2021.04.025

Goldman-Rakic PS (1995) Cellular basis of working memory. Neuron 14:477–485. https://doi.org/10.1016/0896-6273(95)90304-6

Gollisch T, Meister M (2008) Rapid neural coding in the retina with relative spike latencies. Science 319(5866):1108–1111. https://doi.org/10.1126/science.1149639

Graumann M, Ciuffi C, Dwivedi K et al (2022) The spatiotemporal neural dynamics of object location representations in the human brain. Nat Hum Behav 6:796–811. https://doi.org/10.1038/s41562-022-01302-0

Güçlü U, van Gerven MAJ (2015) Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J Neurosci 35:10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Gupta A, Anpalagan A, Guan L, Khwaja AS (2021) Deep learning for object detection and scene perception in self-driving cars: survey, challenges, and open issues. Array 10:100057. https://doi.org/10.1016/j.array.2021.100057

Han J-H, Kushner SA, Yiu AP et al (2009) Selective erasure of a fear memory. Science 323:1492–1496

Haxby JV, Gobbini MI, Furey ML et al (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425–2430. https://doi.org/10.1126/science.1063736

Haxby JV, Connolly AC, Guntupalli JS (2014) Decoding neural representational spaces using multivariate pattern analysis. Annu Rev Neurosci 37:435–456. https://doi.org/10.1146/annurev-neuro-062012-170325

Haynes J-D, Rees G (2006) Decoding mental states from brain activity in humans. Nat Rev Neurosci 7:523–534. https://doi.org/10.1038/nrn1931

Hebart MN, Baker CI (2018) Deconstructing multivariate decoding for the study of brain function. Neuroimage 180:4–18. https://doi.org/10.1016/j.neuroimage.2017.08.005

Hebart MN, Zheng CY, Pereira F, Baker CI (2020) Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nat Hum Behav 4:1173–1185. https://doi.org/10.1038/s41562-020-00951-3

Hebb DO (1949) The Organization of Behavior, Wiley: New York; 1949. Brain research bulletin 50(5-6):437. https://doi.org/10.1016/s0361-9230(99)00182-3

Herten N, Otto T, Wolf OT (2017a) The role of eye fixation in memory enhancement under stress - an eye tracking study. Neurobiol Learn Mem 140:134–144. https://doi.org/10.1016/j.nlm.2017.02.016

Herten N, Pomrehn D, Wolf OT (2017b) Memory for objects and startle responsivity in the immediate aftermath of exposure to the Trier Social Stress Test. Behav Brain Res 326:272–280. https://doi.org/10.1016/j.bbr.2017.03.002

Het S, Ramlow G, Wolf OT (2005) A meta-analytic review of the effects of acute cortisol administration on human memory. Psychoneuroendocrinology 30:771–784. https://doi.org/10.1016/j.psyneuen.2005.03.005

Himmer L, Schönauer M, Heib DPJ et al (2019) Rehearsal initiates systems memory consolidation, sleep makes it last. Sci Adv 5:eaav1695. https://doi.org/10.1126/sciadv.aav1695

Hubel DH, Wiesel TN (1959) Receptive fields of single neurones in the cat's striate cortex. J Physiol 148:574–591. https://doi.org/10.1113/jphysiol.1959.sp006308

Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Huth AG, Nishimoto S, Vu AT, Gallant JL (2012) A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76:1210–1224

Joëls M, Pu Z, Wiegert O et al (2006) Learning under stress: how does it work? Trends Cogn Sci 10:152–158. https://doi.org/10.1016/j.tics.2006.02.002

Josselyn SA, Köhler S, Frankland PW (2015) Finding the engram. Nat Rev Neurosci 16:521–534. https://doi.org/10.1038/nrn4000

Jozwik KM, Kriegeskorte N, Storrs KR, Mur M (2017) Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. Front Psychol 8:1726. https://doi.org/10.3389/fpsyg.2017.01726

Káli S, Dayan P (2004) Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. Nat Neurosci 7:286–294. https://doi.org/10.1038/nn1202

Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. Nat Neurosci 8:679–685. https://doi.org/10.1038/nn1444

Kensinger EA (2007) Negative emotion enhances memory accuracy. Curr Dir Psychol Sci 16(4):213–218. https://doi.org/10.1111/j.1467-8721.2007.00506.x

Kensinger EA, Garoff-Eaton RJ, Schacter DL (2006) Memory for specific visual details can be enhanced by negative arousing content. J Mem Lang 54:99–112. https://doi.org/10.1016/j.jml.2005.05.005

Kensinger EA, Garoff-Eaton RJ, Schacter DL (2007) How negative emotion enhances the visual specificity of a memory. J Cogn Neurosci 19(11):1872–1887. https://doi.org/10.1162/jocn.2007.19.11.1872

Khaligh-Razavi S-M, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol 10:e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Kietzmann TC, McClure P, Kriegeskorte N (2019a) Deep neural networks in computational neuroscience. Oxford research encyclopedia of neuroscience. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264086.013.46

Kietzmann TC, Spoerer CJ, Sörensen LKA et al (2019b) Recurrence is required to capture the representational dynamics of the human visual system. Proc Natl Acad Sci USA 116:21854–21863. https://doi.org/10.1073/pnas.1905544116

Kılıç A, Criss AH, Malmberg KJ, Shiffrin RM (2017) Models that allow us to perceive the world more accurately also allow us to remember past events more accurately via differentiation. Cogn Psychol 92:65–86. https://doi.org/10.1016/j.cogpsych.2016.11.005

Kirschbaum C, Pirke KM, Hellhammer DH (1993) The "Trier Social Stress Test"–a tool for investigating psychobiological stress responses in a laboratory setting. Neuropsychobiology 28:76–81. https://doi.org/10.1159/000119004

Konkle T, Brady TF, Alvarez GA, Oliva A (2010) Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. J Exp Psychol Gen 139:558–578. https://doi.org/10.1037/a0019165

Kragel PA, LaBar KS (2016) Decoding the nature of emotion in the brain. Trends Cogn Sci 20:444–455. https://doi.org/10.1016/j.tics.2016.03.011

Kragel PA, Koban L, Barrett LF, Wager TD (2018) Representation, pattern information, and brain signatures: from neurons to neuroimaging. Neuron 99:257–273. https://doi.org/10.1016/j.neuron.2018.06.009

Kragel PA, Reddan MC, LaBar KS, Wager TD (2019) Emotion schemas are embedded in the human visual system. Sci Adv 5(7):eaaw4358. https://doi.org/10.1126/sciadv.aaw4358

Kravitz DJ, Saleem KS, Baker CI et al (2013) The ventral visual pathway: an expanded neural framework for the processing of object quality. Trends Cogn Sci 17:26–49. https://doi.org/10.1016/j.tics.2012.10.011

Kriegeskorte N, Diedrichsen J (2019) Peeling the onion of brain representations. Annu Rev Neurosci 42:407–432. https://doi.org/10.1146/annurev-neuro-080317-061906

Kriegeskorte N, Golan T (2019) Neural network models and deep learning. Curr Biol 29:R231–R236. https://doi.org/10.1016/j.cub.2019.02.034

Kriegeskorte N, Kievit RA (2013) Representational geometry: integrating cognition, computation, and the brain. Trends Cogn Sci 17:401–412. https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte N, Wei X-X (2021) Neural tuning and representational geometry. Nat Rev Neurosci 22:703–718. https://doi.org/10.1038/s41583-021-00502-3

Kriegeskorte N, Mur M, Bandettini P (2008a) Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci 2:4. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte N, Mur M, Ruff DA et al (2008b) Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60:1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60:84–90. https://doi.org/10.1145/3065386

Kubilius J, Schrimpf M, Nayebi A et al (2018) CORnet: modeling the neural mechanisms of core object recognition. bioRxiv https://doi.org/10.1101/408385

Kubilius J, Schrimpf M, Kar K et al (2019) Brain-like object recognition with high-performing shallow recurrent ANNs. Adv Neural Inform Process Syst. arXiv:1909.06161

Kumaran D, Hassabis D, McClelland JL (2016) What learning systems do intelligent agents need? Complementary learning systems theory updated. Trends Cogn Sci 20:512–534

Kunz L, Deuker L, Zhang H, Axmacher N (2018) Tracking human engrams using multivariate analysis techniques. Handbook of behavioral neuroscience. Elsevier, pp 481–508. https://doi.org/10.1016/B978-0-12-812028-6.00026-4

Kuzovkin I, Vicente R, Petton M et al (2018) Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. Commun Biol 1:107. https://doi.org/10.1038/s42003-018-0110-y

LaBar KS, Cabeza R (2006) Cognitive neuroscience of emotional memory. Nat Rev Neurosci 7:54–64. https://doi.org/10.1038/nrn1825

LaRocque KF, Smith ME, Carr VA et al (2013) Global similarity and pattern separation in the human medial temporal lobe predict subsequent memory. J Neurosci 33:5466–5474. https://doi.org/10.1523/JNEUROSCI.4293-12.2013

Lashley K (1950) In search of the engram. Symp Soc Exp Biol 4:454–482

Lee H, Chen J (2022) Predicting memory from the network structure of naturalistic events. Nat Commun 13:4235. https://doi.org/10.1038/s41467-022-31965-2

Lee H, Kuhl BA (2016) Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex. J Neurosci 36:6069–6082. https://doi.org/10.1523/JNEUROSCI.4286-15.2016

Leeds DD, Seibert DA, Pyles JA, Tarr MJ (2013) Comparing visual representations across human fMRI and computational vision. J vis 13:25. https://doi.org/10.1167/13.13.25

Lewis PA, Durrant SJ (2011) Overlapping memory replay during sleep builds cognitive schemata. Trends Cogn Sci 15:343–351. https://doi.org/10.1016/j.tics.2011.06.004

Lifanov J, Linde-Domingo J, Wimber M (2021) Feature-specific reaction times reveal a semanticisation of memories over time and with repeated remembering. Nat Commun 12:3177. https://doi.org/10.1038/s41467-021-23288-5

Lin B, Mur M, Kietzmann T, Kriegeskorte N (2019) Visualizing representational dynamics with multidimensional scaling alignment. arXiv:1906.09264

Linde-Domingo J, Treder MS, Kerrén C, Wimber M (2019) Evidence that neural information flow is reversed between object perception and object reconstruction from memory. Nat Commun 10:179. https://doi.org/10.1038/s41467-018-08080-2

Liu X, Ramirez S, Pang PT et al (2012) Optogenetic stimulation of a hippocampal engram activates fear memory recall. Nature 484:381–385. https://doi.org/10.1038/nature11028

Liu X, Ramirez S, Tonegawa S (2014a) Inception of a false memory by optogenetic manipulation of a hippocampal memory engram. Philos Trans R Soc Lond B Biol Sci 369:20130142. https://doi.org/10.1098/rstb.2013.0142

Liu X, Ramirez S, Redondo RL, Tonegawa S (2014b) Identification and manipulation of memory engram cells. Cold Spring Harb Symp Quant Biol 79:59–65. https://doi.org/10.1101/sqb.2014.79.024901

Liu Y, Dolan RJ, Kurth-Nelson Z, Behrens TEJ (2019) Human replay spontaneously reorganizes experience. Cell 178:640-652.e14. https://doi.org/10.1016/j.cell.2019.06.012

Liu J, Zhang H, Yu T et al (2020) Stable maintenance of multiple representational formats in human visual short-term memory. Proc Natl Acad Sci USA 117:32329–32339. https://doi.org/10.1073/pnas.2006752117

Liu J, Zhang H, Yu T et al (2021) Transformative neural representations support long-term episodic memory. Sci Adv 7:eabg9715. https://doi.org/10.1126/sciadv.abg9715

Luck SJ, Vogel EK (1997) The capacity of visual working memory for features and conjunctions. Nature 390:279–281. https://doi.org/10.1038/36846

Macé MJ-M, Joubert OR, Nespoulous J-L, Fabre-Thorpe M (2009) The time-course of visual categorizations: you spot the animal faster than the bird. PLoS One 4:e5927. https://doi.org/10.1371/journal.pone.0005927

Marblestone AH, Wayne G, Kording KP (2016) Toward an integration of deep learning and neuroscience. Front Comput Neurosci 10:94. https://doi.org/10.3389/fncom.2016.00094

Martin CB, Douglas D, Newsome RN et al (2018) Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. Elife. https://doi.org/10.7554/eLife.31873

Mather M (2007) Emotional arousal and memory binding: an object-based framework. Perspect Psychol Sci 2:33–52. https://doi.org/10.1111/j.1745-6916.2007.00028.x

McClure P, Kriegeskorte N (2016) Representational distance learning for deep neural networks. Front Comput Neurosci 10:131. https://doi.org/10.3389/fncom.2016.00131

McGrath T, Kapishnikov A, Tomašev N et al (2022) Acquisition of chess knowledge in AlphaZero. Proc Natl Acad Sci USA 119:e2206625119. https://doi.org/10.1073/pnas.2206625119

Meyers EM, Freedman DJ, Kreiman G et al (2008) Dynamic population coding of category information in inferior temporal and prefrontal cortex. J Neurophysiol 100:1407–1419. https://doi.org/10.1152/jn.90248.2008

Mur M, Bandettini PA, Kriegeskorte N (2009) Revealing representational content with pattern-information fMRI–an introductory guide. Soc Cogn Affect Neurosci 4:101–109. https://doi.org/10.1093/scan/nsn044

Mur M, Meys M, Bodurka J et al (2013) Human object-similarity judgments reflect and transcend the primate-IT object representation. Front Psychol 4:128. https://doi.org/10.3389/fpsyg.2013.00128

Naselaris T, Kay KN, Nishimoto S, Gallant JL (2011) Encoding and decoding in fMRI. Neuroimage 56:400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Naspi L, Hoffman P, Devereux B et al (2021a) Multiple dimensions of semantic and perceptual similarity contribute to mnemonic discrimination for pictures. J Exp Psychol Learn Mem Cogn 47:1903–1923. https://doi.org/10.1037/xlm0001032

Naspi L, Hoffman P, Devereux B, Morcom AM (2021b) Perceptual and semantic representations at encoding contribute to true and false recognition of objects. J Neurosci 41:8375–8389. https://doi.org/10.1523/JNEUROSCI.0677-21.2021

Newen A, Vosgerau G (2020) Situated mental representations. What are mental representations? Oxford University Press, pp 178–212

Nonaka S, Majima K, Aoki SC, Kamitani Y (2021) Brain hierarchy score: Which deep neural networks are hierarchically brain-like? iScience 24:103013. https://doi.org/10.1016/j.isci.2021.103013

O'Donnell RE, Clement A, Brockmole JR (2018) Semantic and functional relationships among objects increase the capacity of visual working memory. J Exp Psychol Learn Mem Cogn 44:1151–1158. https://doi.org/10.1037/xlm0000508

O'Reilly RC, Bhattacharyya R, Howard MD, Ketz N (2014) Complementary learning systems. Cogn Sci 38:1229–1248. https://doi.org/10.1111/j.1551-6709.2011.01214.x

Paller KA, Wagner AD (2002) Observing the transformation of experience into memory. Trends Cogn Sci 6:93–102. https://doi.org/10.1016/s1364-6613(00)01845-3

Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45:S199-209. https://doi.org/10.1016/j.neuroimage.2008.11.007

Perolat J, De Vylder B, Hennes D et al (2022) Mastering the game of Stratego with model-free multiagent reinforcement learning. Science 378:990–996. https://doi.org/10.1126/science.add4679

Poldrack RA (2021) The physics of representation. Synthese 199:1307–1325. https://doi.org/10.1007/s11229-020-02793-y

Popel M, Tomkova M, Tomek J, Kaiser L, Uszkoreit J, Bojar O, Žabokrtský Z (2020) Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. Nat Commun 11(1):4381. https://doi.org/10.1038/s41467-020-18073-9

Raizada RDS, Tsao F-M, Liu H-M, Kuhl PK (2010) Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: prediction of individual differences. Cereb Cortex 20:1–12. https://doi.org/10.1093/cercor/bhp076

Randall B, Moss HE, Rodd JM et al (2004) Distinctiveness and correlation in conceptual structure: behavioral and computational studies. J Exp Psychol Learn Mem Cogn 30:393–406. https://doi.org/10.1037/0278-7393.30.2.393

Ranganath C (2010) Binding items and contexts. Curr Dir Psychol Sci 19:131–137. https://doi.org/10.1177/0963721410368805

Reddy L, Kanwisher N (2006) Coding of visual objects in the ventral stream. Curr Opin Neurobiol 16(4):408–414. https://doi.org/10.1016/j.conb.2006.06.004

Richards BA, Lillicrap TP, Beaudoin P et al (2019) A deep learning framework for neuroscience. Nat Neurosci 22:1761–1770. https://doi.org/10.1038/s41593-019-0520-2

Roozendaal B (2002) Stress and memory: opposing effects of glucocorticoids on memory consolidation and memory retrieval. Neurobiol Learn Mem 78:578–595. https://doi.org/10.1006/nlme.2002.4080

Roskies AL (2021) Representational similarity analysis in neuroimaging: proxy vehicles and provisional representations. Synthese 199:5917–5935. https://doi.org/10.1007/s11229-021-03052-4

Saxe A, Nelli S, Summerfield C (2021) If deep learning is the answer, what is the question? Nat Rev Neurosci 22:55–67. https://doi.org/10.1038/s41583-020-00395-8

Schacter DL (2001) Forgotten ideas, neglected pioneers: richard semon and the story of memory. Psychology Press. https://doi.org/10.4324/9780203720134

Schacter DL, Addis DR (2007) The cognitive neuroscience of constructive memory: remembering the past and imagining the future. Philos Trans R Soc Lond B Biol Sci 362:773–786. https://doi.org/10.1098/rstb.2007.2087

Schapiro AC, McDevitt EA, Chen L et al (2017) Sleep benefits memory for semantic category structure while preserving exemplar-specific information. Sci Rep 7:14869. https://doi.org/10.1038/s41598-017-12884-5

Semon R (1904) Die Mneme als erhaltendes Prinzip im Wechsel des organischen Geschehens. Wilhelm Engelmann, Leipzig

Semon R (1909) Die nmemischen Empfindungen. Wilhelm Engelmann, Leipzig

Shea N (2018) Representation in cognitive science. Oxford University Press, London, England

Shields GS, Sazma MA, McCullough AM, Yonelinas AP (2017) The effects of acute stress on episodic memory: a meta-analysis and integrative review. Psychol Bull 143:636–675. https://doi.org/10.1037/bul0000100

Silver D, Schrittwieser J, Simonyan K et al (2017) Mastering the game of go without human knowledge. Nature 550:354–359. https://doi.org/10.1038/nature24270

Skinner BF (1953) Science and human behavior. Macmillan, London/New York City

Stokes MG (2015) "Activity-silent" working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn Sci 19:394–405. https://doi.org/10.1016/j.tics.2015.05.004

Storrs KR, Kriegeskorte N (2019) Deep learning for cognitive neuroscience. arXiv:1903.01458

Talmi D (2013) Enhanced emotional memory. Curr Dir Psychol Sci 22:430–436. https://doi.org/10.1177/0963721413498893

Taylor KI, Devereux BJ, Acres K et al (2012) Contrasting effects of feature-based statistics on the categorisation and basic-level identification of visual objects. Cognition 122:363–374. https://doi.org/10.1016/j.cognition.2011.11.001

Ten Oever S, Sack AT, Oehrn CR, Axmacher N (2021) An engram of intentionally forgotten information. Nat Commun 12:6443. https://doi.org/10.1038/s41467-021-26713-x

Tolman EC (1948) Cognitive maps in rats and men. Psychol Rev 55(4):189–208. https://doi.org/10.4324/9780203789155-11

Tulving E, Watkins MJ (1975) Structure of memory traces. Psychol Rev 82:261–275

Tyler LK, Chiu S, Zhuang J et al (2013) Objects and categories: feature statistics and object processing in the ventral stream. J Cogn Neurosci 25:1723–1735. https://doi.org/10.1162/jocn_a_00419

van Bergen RS, Kriegeskorte N (2020) Going in circles is the way forward: the role of recurrence in visual inference. Curr Opin Neurobiol 65:176–193. https://doi.org/10.1016/j.conb.2020.11.009

van Kesteren MTR, Beul SF, Takashima A et al (2013) Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: from congruent to incongruent. Neuropsychologia 51:2352–2359. https://doi.org/10.1016/j.neuropsychologia.2013.05.027

Vilarroya O (2017) Neural representation. A survey-based analysis of the notion. Front Psychol 8:1458. https://doi.org/10.3389/fpsyg.2017.01458

Vinyals O, Babuschkin I, Czarnecki WM et al (2019) Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature 575:350–354. https://doi.org/10.1038/s41586-019-1724-z

Visser RM, Scholte HS, Kindt M (2011) Associative learning increases trial-by-trial similarity of BOLD-MRI patterns. J Neurosci 31:12021–12028. https://doi.org/10.1523/JNEUROSCI.2178-11.2011

Visser RM, Scholte HS, Beemsterboer T, Kindt M (2013) Neural pattern similarity predicts long-term fear memory. Nat Neurosci 16:388–390. https://doi.org/10.1038/nn.3345

Vo VA, Sutterer DW, Foster JJ et al (2022) Shared representational formats for information maintained in working memory and information retrieved from long-term memory. Cereb Cortex 32:1077–1092. https://doi.org/10.1093/cercor/bhab267

Watrous AJ, Fell J, Ekstrom AD, Axmacher N (2015) More than spikes: common oscillatory mechanisms for content specific neural representations during perception and memory. Curr Opin Neurobiol. https://doi.org/10.1016/j.conb.2014.07.024

Watson JB (1913) Psychology as the behaviorist views it. Psychol Rev 20:158–177. https://doi.org/10.1037/h0074428

Wen H, Shi J, Zhang Y et al (2018) Neural encoding and decoding with deep learning for dynamic natural vision. Cereb Cortex 28:4136–4160. https://doi.org/10.1093/cercor/bhx268

Wiemers US, Schoofs D, Wolf OT (2013) A friendly version of the trier social stress test does not activate the HPA axis in healthy men and women. Stress. https://doi.org/10.3109/10253890.2012.714427

Wiemers US, Sauvage MM, Wolf OT (2014) Odors as effective retrieval cues for stressful episodes. Neurobiol Learn Mem 112:230–236. https://doi.org/10.1016/j.nlm.2013.10.004

Wing EA, Geib BR, Wang W-C et al (2020) Cortical overlap and cortical-hippocampal interactions predict subsequent true and false memory. J Neurosci 40:1920–1930. https://doi.org/10.1523/JNEUROSCI.1766-19.2020

Winocur G, Moscovitch M (2011) Memory transformation and systems consolidation. J Int Neuropsychol Soc 17:766–780. https://doi.org/10.1017/S1355617711000683

Wolf OT (2009) Stress and memory in humans: twelve years of progress? Brain Res 1293:142–154. https://doi.org/10.1016/j.brainres.2009.04.013

Wolf OT (2019) Memories of and influenced by the Trier Social Stress Test. Psychoneuroendocrinology 105:98–104. https://doi.org/10.1016/j.psyneuen.2018.10.031

Xue G (2018) The neural representations underlying human episodic memory. Trends Cogn Sci 22:544–561. https://doi.org/10.1016/j.tics.2018.03.004

Xue G (2022) From remembering to reconstruction: the transformative neural representation of episodic memory. Prog Neurobiol 219:102351. https://doi.org/10.1016/j.pneurobio.2022.102351

Xue G, Dong Q, Chen C et al (2010) Greater neural pattern similarity across repetitions is associated with better memory. Science 330:97–101. https://doi.org/10.1126/science.1193125

Yamins DLK, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci 19:356–365. https://doi.org/10.1038/nn.4244

Yassa MA, Stark CEL (2011) Pattern separation in the hippocampus. Trends Neurosci 34:515–525. https://doi.org/10.1016/j.tins.2011.06.006

Yonelinas AP, Ritchey M (2015) The slow forgetting of emotional episodic memories: an emotional binding account. Trends Cogn Sci 19:259–267. https://doi.org/10.1016/j.tics.2015.02.009

Zhou B, Khosla A, Lapedriza A et al (2015) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929. arXiv:1512.04150

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.