Mini-Symposium

Brain Mechanisms of Concept Learning

Dagmar Zeithamova,[1]* Michael L. Mack,[2]* Kurt Braunlich,[3] Tyler Davis,[4] Carol A. Seger,[5,3] Marlieke T.R. van Kesteren,[6,9] and Andreas Wutz[7,8]

[1]Department of Psychology and Institute of Neuroscience, University of Oregon, Eugene, Oregon 97403, [2]Department of Psychology, University of Toronto, Toronto, Ontario M5S 3G3, Canada, [3]Department of Psychology and Program in Molecular, Cellular, and Integrative Neurosciences, Colorado State University, Fort Collins, Colorado 80523, [4]Department of Psychological Sciences, Texas Tech University, Lubbock, Texas 79403, [5]Center for the Study of Applied Psychology, Key Laboratory of Mental Health and Cognitive Science of Guangdong Province, School of Psychology, South China Normal University, Guangzhou 510631, China, [6]Section of Education Sciences and LEARN! Research Institute, Vrije Universiteit Amsterdam, Amsterdam 1081 BT, The Netherlands, [7]The Picower Institute for Learning & Memory and Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, [8]Center for Cognitive Neuroscience, University of Salzburg, Hellbrunnerstrasse 34, 5020 Salzburg, Austria, and [9]Institute for Brain and Behavior Amsterdam, Vrije Universiteit Amsterdam, Amsterdam 1081 BT, The Netherlands

Concept learning, the ability to extract commonalities and highlight distinctions across a set of related experiences to build organized knowledge, is a critical aspect of cognition. Previous reviews have focused on concept learning research as a means for dissociating multiple brain systems. The current review surveys recent work that uses novel analytical approaches, including the combination of computational modeling with neural measures, focused on testing theories of specific computations and representations that contribute to concept learning. We discuss in detail the roles of the hippocampus, ventromedial prefrontal, lateral prefrontal, and lateral parietal cortices, and how their engagement is modulated by the coherence of experiences and the current learning goals. We conclude that the interaction of multiple brain systems relating to learning, memory, attention, perception, and reward support a flexible concept-learning mechanism that adapts to a range of category structures and incorporates motivational states, making concept learning a fruitful research domain for understanding the neural dynamics underlying complex behaviors.

Key words: **categorization; computational modeling; hippocampus; prefrontal cortex; parietal cortex; fMRI**

Introduction

How do we build new concepts from our experiences to support generalization to novel situations? How do we categorize a new animal as a dog given previous experiences with dogs (Fig. 1)? Do we follow rules and definitions or base decisions on similarity (Bruner et al., 1967; Rosch and Mervis, 1975)? Do we retrieve representations of specific dogs we encountered or have we formed an abstract concept of a dog that transcends specific experiences (Posner and Keele, 1968; Nosofsky, 1986)?

Over the last few decades, a wide range of cognitive processes and brain regions have been implicated in concept learning, including those related to memory, reasoning, decision-making, and reward processing. Early neuroscience work on category learning focused on establishing how different modes of category acquisition involved dissociable brain systems. For example, different types of categories may be acquired through memorization of individual category examples by the episodic memory system, hypothesis-testing, and rule-abstraction supported by the lateral prefrontal cortex, or incremental learning processes such as perceptual learning or feedback-based procedural learning supported by non-declarative memory systems (Aizenstein et al., 2000; Nomura et al., 2007; Casale and Ashby, 2008; Zeithamova et al., 2008; Ashby and Maddox, 2011; Morrison et al., 2015). Such system-level dissociations have been the focus of several previous reviews of category learning research (Ashby and Maddox, 2005, 2011; Seger and Miller, 2010).

Current neuroscience research has begun to move past the system-level dissociations toward developing computational theories to test specific candidate mechanisms for brain regions involved in category learning. This new computational revolution has focused heavily on two neurobiological divisions: the medial temporal lobes and the prefrontal cortex. Work on the medial temporal lobes has focused on how functions ascribed to the hippocampus, such as pattern completion, pattern separation, and memory integration, can be applied to concept learning and categorization (Mack et al., 2018). Although the initial patient work suggested a limited role for the hippocampus in concept learning (Knowlton and Squire, 1993; Squire and Knowlton, 1995), by combining existing formal models of learning theories (Posner and Keele, 1968; Nosofsky, 1986; Love et al., 2004) with

Received June 21, 2019; revised Aug. 6, 2019; accepted Aug. 9, 2019.

This work was supported in part by NSERC Discovery Grant RGPIN-2017-06753 (M.L.M.), Lewis Family Endowment that supports the Lewis Center for Neuroimaging at the University of Oregon (D.Z.), Chang Jiang Scholars program of the Ministry of Education, China (C.A.S.), Marie Curie Individual Fellowship of the EU Horizon 2020 Framework Program for Research and Innovation 704506 (M.T.R.v.K.), National Institutes of Health R01MH065252 and R37MH087027 (A.W.), MIT Picower Institute Innovation Fund (A.W.), and Austrian Science Fund M02496 (A.W.).

The authors declare no competing financial interests.

*D.Z. and M.L.M. contributed equally to this work; the remaining authors are listed in an alphabetical order.

Correspondence should be addressed to Dagmar Zeithamova at dasa@uoregon.edu or Michael L. Mack at mack@psych.utoronto.ca.

https://doi.org/10.1523/JNEUROSCI.1166-19.2019

human neuroimaging, recent work posits a more central role (Davis et al., 2012a,b; Mack et al., 2016; Schapiro et al., 2017; Bowman and Zeithamova, 2018). Within the prefrontal cortex, ventromedial prefrontal cortex has been of interest in categorization because of its involvement in other forms of generalization, such as representation of schemas (van Kesteren et al., 2012; Gilboa and Marlatte, 2017) and linking related memories into an integrated representation (Schlichting and Preston, 2015). Hierarchical control theory, suggesting a possible rostrocaudal gradient of representational abstraction across the lateral prefrontal cortex (Badre and D'esposito, 2009; Badre and Nee, 2018), has inspired recent work on characterizing how distinct category learning contexts differentially engage specific cognitive control mechanisms and reasoning processes supported by subregions of lateral PFC (Paniukov and Davis, 2018). Finally, computational neuroimaging approaches have identified processes and representations contributed by the lateral parietal cortex, a region that has not played a major role in systems-level dissociations, but importantly contains both stimulus-specific and categorical memory representations (Freedman and Assad, 2006, 2016; Kuhl and Chun, 2014; Sestieri et al., 2017). The present review covers these recent developments, highlighting how innovative approaches that bridge computational modeling and neural measures provide novel insights into the computations and representations involved in concept learning.

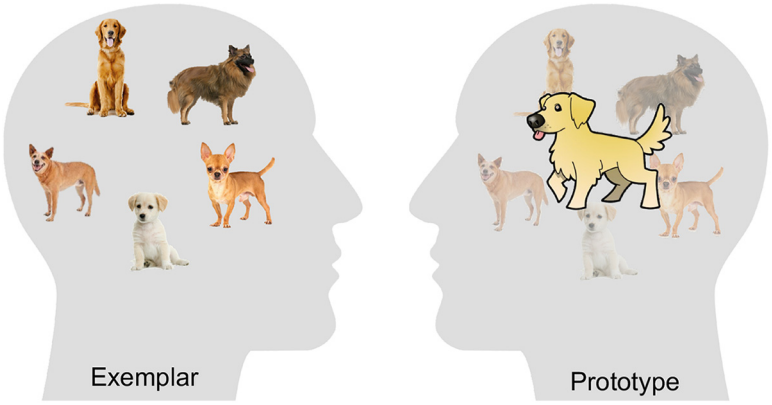

Figure 1. Concept representations per the exemplar and prototype models. Exemplar model assumes that categories are represented by specific exemplars. Prototype model assumes that categories are represented by their central tendency (prototype), which is abstracted from specific exemplars and embodies all characteristic features.

The dynamic formation of concept representations during category learning

Concept learning is rapid and flexible; we can adapt newly-learned knowledge to novel situations or changing goals with seemingly little effort. Characterizing the neural mechanisms that support such rapid conceptual learning, especially learning in the face of evolving learning goals, is a critical focus of ongoing research. One recent study (Mack et al., 2016) addressed this topic with an approach combining fMRI and computational modeling to identify the neural machinery of new concept formation. Motivated by human and animal work demonstrating attentional and contextual influences on representations in the hippocampus (Fenton et al., 2010; Aly and Turk-Browne, 2016), this study targeted the formation of new concepts in hippocampal activation patterns. Across two learning tasks, participants learned to categorize multidimensional visual objects according to different category structures: a simple rule based only on a single feature dimension and a more complicated rule based on two feature dimensions. Critically, the visual objects were constant across the tasks, but the underlying organization of the objects into categories differed. This approach required participants to attend to different features between tasks and build new conceptual representations that best matched the underlying category structure.

To investigate the nature of the representations each participant learned in the two tasks, Mack et al. (2016) used SUSTAIN, a computational model that accounts well for behavioral responses during category learning (Love et al., 2004; Love and Gureckis, 2007). Fitting this model separately to each participant's learning behavior provided predictions about how category items were represented in a multidimensional space, and thus how similar they may be perceived by the participant, across the two tasks. The logic followed such that the more similar any two items were represented by the model, the more similar should be their neural activation patterns. To test the role of the hippocampus in building flexible concept knowledge, the model-based predictions were used to interrogate the structure of activation patterns in the hippocampus with representational similarity analysis (Kriegeskorte et al., 2008). If the hippocampus plays a role in building flexible concept knowledge, SUSTAIN′s predictions of category structure should be evident in hippocampal representations. Indeed, a region in anterior hippocampus showed neural representations that were consistent with the model-based similarity matrices. In other words, as learning goals changed, hippocampal representations reorganized to reflect the diagnostic information important for the current task.

Although hippocampal functions offer a powerful toolset for building new concepts (Kumaran et al., 2009; Schapiro et al., 2017), the hippocampus does not act alone. In particular, both rodent (Place et al., 2016; Guise and Shapiro, 2017) and human (van Kesteren et al., 2012; Zeithamova et al., 2012; Schlichting and Preston, 2016) findings point to a functional alliance between hippocampus and medial PFC when encoding new information that overlaps with prior experiences (Preston and Eichenbaum, 2013). Consistent with this notion, Mack et al. (2016) found that anterior hippocampus demonstrated a strong functional coupling with ventromedial PFC (vmPFC) during early learning when concept updating is most potent. This region of vmPFC may, in fact, play an important role itself in guiding attentional tuning during new learning. Neural representations in vmPFC throughout new concept learning have recently been shown to systematically track the efficient mapping of stimuli to categories in a manner consistent with highlighting information that matters and down-weighting irrelevant features (Mack et al., 2019).

Specific and generalized representations supporting categorization

What type or types of memory representations are formed during concept learning and used in categorization decisions has been the focus of cognitive categorization research for decades (Posner and Keele, 1968; Homa, 1973; Medin and Schaffer, 1978; Nosofsky, 1986). Combining neuroimaging with formal models of categorization made it possible to start answering such representational questions in neuroscience research. A study by Mack et

al. (2013) tested the degree to which brain signals are consistent with the exemplar model (assuming that categories are represented by the specific category examples one encountered) and the prototype model of categorization (assuming that categories are represented by generalized category representations, or prototypes, abstracted across exemplars). These models are schematically depicted in Figure 1. They identified a network of regions representing specific exemplars during generalization judgments but found no evidence of generalized prototype representations. The categories tested by Mack et al. (2013) had relatively low coherence, where some stimuli were equally distant from the central tendency of their category and that of the other category. Bowman and Zeithamova (2018) used a similar model-based fMRI methodology, but a different category structure that included greater coherence of stimuli within each category. They found evidence consistent with the predictions of the prototype model in both behavior and the brain. Bowman and Zeithamova (2019) then tested the idea that concept learning may involve both specific (exemplar) and generalized (prototype) representations under different conditions and found that coherence of the underlying category structure is a critical factor determining the success of forming a generalized category representation (Minda and Smith, 2001).

Theoretically, the same network of regions could be involved in concept representation across different category structures, with only the nature of the representations formed being different. However, recent findings are more indicative of distinct neural mechanisms existing for representing specific instances versus summary representations, with these mechanisms being differentially engaged for different category structures. Mack et al. (2013) identified a network of regions representing specific exemplars that included lateral prefrontal and parietal regions implicated in supporting memory for specific events and preventing interference in episodic memory tasks (Badre and Wagner, 2005; Kuhl and Chun, 2014). Regions tracking prototype model predictions by Bowman and Zeithamova (2018) included hippocampus and ventromedial prefrontal cortex. Although no above-threshold evidence for exemplar representation during concept generalization was found in that study, subthreshold exemplar regions were consistent with those found by Mack et al. (2013) and distinct from those tracking prototype predictors. Thus, distinct neural mechanisms creating different types of may be engaged in concept learning, with their contribution varying across category structure. We will discuss further evidence for this notion from nonhuman primate research later in this review (Wutz et al., 2018).

The notion that both specific and generalized memories may represent concepts and support generalization aligns with findings from a different generalization task: episodic inference. In a typical task, participants learn a set of overlapping associations, such as A relates to B and B relates to C (Underwood, 1949). Although many behavioral and neuroimaging studies showed that reactivation of A during BC-learning can lead to interference and forgetting of C (Anderson, 2003; Kuhl et al., 2011), another outcome is episodic inference, where a relation is inferred between A and C. Although episodic inference can be achieved on-demand, from separate memories of individual events (Kumaran and McClelland, 2012), it can also result from memory integration, whereby new events are linked with prior related memories into a combined representation (Schlichting et al., 2014; Richter et al., 2016; Zeithamova and Preston, 2017). Notably, the same hippocampal-vmPFC interactions implicated in concept learning (Kumaran et al., 2009; Bowman and Zeithamova, 2018; Frank et al., 2019) have been implicated in memory integration and inference in both neuroimaging (Zeithamova et al., 2012; Schlichting et al., 2015) and lesion work (Dusek and Eichenbaum, 1997; Ryan et al., 2016; Spalding et al., 2018). Finally, the same regions have been shown to underlie schema-related memory (Tse et al., 2007, 2011; van Kesteren et al., 2012; Spalding et al., 2015; Brod et al., 2017; Gilboa and Marlatte, 2017; Gilboa and Moscovitch, 2017; Baldassano et al., 2018; Romero et al., 2019). Thus, hippocampal-vmPFC memory integration mechanisms may serve to link related information to a coherent representation in service of a range of generalization tasks, including concept learning.

Congruency and reactivation aid memory integration

Understanding how the brain links related information may help us to take a step further and focus on how memory integration processes can be enhanced, as desired, for example, in educational settings. Factors such as congruency between associates and reactivation of previously learned information have been shown to facilitate memory integration (Van Kesteren et al., 2018). However, how the brain achieves such memory improvements and how they link to memory integration processes is yet unknown.

To examine how enhanced memory integration is realized, van Kesteren et al. (2019) extended the AB-BC inference paradigm described in the previous section. They added a factor of congruency between A and C (a scene and an object), which is a strong enhancing factor in conventional associative memory experiments (van Kesteren et al., 2012, 2013a,b), and asked participants to rate the strength of reactivation (subjective reactivation of the scene) they experienced during BC-learning (pseudoword with object). Both these congruency and reactivation factors were expected to yield improved memory integration as was found in a recent experiment that used a more educationally valid version of this paradigm (van Kesteren et al., 2018).

This behavioral effect was replicated, and brain activity during memory integration was correlated with three behavioral factors: memory, congruency, and reactivation. MTL, including the hippocampus, was related to memory performance, whereas congruency was associated with activity in the vmPFC and the hippocampus, and reactivation strength revealed an extensive retrieval network that included the vmPFC and the hippocampus. Thus, hippocampus-vmPFC mechanisms may contribute to several aspects and modulatory factors involved in concept learning. This work extends prior work on schema-related memory benefits (Ghosh and Gilboa, 2014; Ryan et al., 2016) and provides new insights into how schema-congruency and active reactivation of existing knowledge improves memory and integration.

Integration of reward and concept representations in categorization

Although we so far discussed the role of vmPFC in memory integration, schema formation, and the acquisition of conceptual representations, a separate line of research on economic decision making emphasizes the role of vmPFC in reward processing (Kable and Glimcher, 2009). Reward is an important factor affecting categorization (Seger and Peterson, 2013), but has so far received little attention in neuroscientific studies of concept learning. Intriguingly, research into reward and value-based decision making has identified cortical regions that are also important for categorization (Summerfield and Tsetsos, 2012; Jocham et al., 2014). Here we focus on the vmPFC and an additional area, the intraparietal sulcus (IPS)/inferior parietal (IP), both of which

are associated with both reward and concept representation. IPS/IP is active during categorization and is sensitive to category representation in humans (Seger et al., 2015; Wheeler et al., 2015) and nonhuman primates (Sarma et al., 2016).

Braunlich and Seger (2016; K. Braunlich & C.A. Seger, unpublished observations) examined how IPS/IP and vmPFC are involved in processing the sum of evidence for category membership and associated reward. Braunlich and Seger (2016) developed a task in which four features probabilistically related to category membership were presented in series over time. Bayesian model selection showed that the interaction between evidence and time to end of trial better accounted for activity than evidence alone in IPS/IP. K. Braunlich and C.A. Seger (unpublished observations) examined how reward availability affected evidence integration and decision during categorization using this task. Subjects were informed about when reward would be available; for example, correct responses would yield a reward when the third feature was presented, but not before or after. Activity in IPS/IP increased as a function of the amount of evidence and time to reward. In addition, the vmPFC and anterior hippocampus showed a ramping pattern of activity as time of reward approached, indicating representation of both the sum of evidence and the temporal distance for reward availability. Functional connectivity analysis found hippocampus connectivity with visual cortex, and vmPFC with both visual cortex and frontoparietal regions, indicating that the vmPFC may receive input from these regions that can be used to monitor the ongoing decision process and maximize reward.

Recent complementary studies have examined how reward expectation is integrated with representational knowledge during categorization. Braunlich et al. (2017) had subjects learn to categorize abstract polygonal stimuli formed as distortions of a prototypical stimulus according to two different decision criteria and found that both IPS/IP and vmPFC were sensitive to distance in perceptual space from the stimulus to the current criterion. C.A. Seger, K. Braunlich, and Z. Liu (unpublished observations) required subjects to combine categorization with information about reward probability to predict outcomes. On each trial, participants saw a cue indicating the probability that correct performance would be rewarded (0, 25, 50, 75, or 100%), then viewed and categorized a stimulus, and finally received reward. During stimulus presentation, IPS/IP was sensitive to the interaction of prototype distance and reward probability, and coded for reward, category, and their interaction in a representational similarity analysis. IPS/IP was also sensitive to both reward and prototype distance prediction error at the time of feedback, indicating a possible role in updating representations based on feedback. The vmPFC was sensitive to reward probability at cue and feedback, indicating a role in maintaining contextual information about reward probability across the trial and integrating it with categorical information. This work provides an intersection of vmPFC research on memory integration discussed here with otherwise separate research that emphasizes the role of vmPFC in subjective value representation through integration of multiple converging inputs (Clithero and Rangel, 2014; Berkman et al., 2017).

Modeling the role of the rostrolateral prefrontal cortex in category learning and generalization

Most concept-learning studies and general computational literature have focused on memory processes and similarity-based mechanisms: how memory representations are built around common information and retrieved on the basis of representational overlap to support novel judgments. Similarity-based processes are critical for long-term category representation (Nosofsky, 1986), however, many theories suggest that such processes are supplemented, in some cases, by inferential processes that are akin to using logical rules or reasoning strategies (Smith and Sloman, 1994; Ashby et al., 1998). For example, rules may augment categorization during early stages of learning while long-term representations are being formed and may be called upon again during generalization when people are confronted with stimuli that are ambiguous or similar to multiple previously acquired category representations (Nosofsky et al., 1994; Palmeri and Nosofsky, 1995; Erickson and Kruschke, 1998; Juslin et al., 2001).

How inferential categorization processes are instantiated in the brain is an open question. Broad neurobiological category-learning theories have focused on the lateral PFC as a whole in the operation of inferential and executive reasoning processes (Seger and Miller, 2010; Ashby and Maddox, 2011). However, they do not attempt to distinguish among subregions of the lateral PFC with respect to such processes. In the broader literature, abstract reasoning processes such as higher-level reasoning, problem solving, and analogy are often thought to depend on the rostrolateral PFC (rlPFC; Christoff et al., 2001; Kroger et al., 2002; Bunge et al., 2005; Green et al., 2006; Hampshire et al., 2011; Watson and Chatterjee, 2012). Evidence for this comes not only from activation-based studies of such abstract reasoning tasks, but also from anatomical data, such as the area's relative size in humans compared with other primates (Burgess et al., 2007) and its growth trajectory during development, which tracks the emergence of higher-level reasoning in children and adolescence (Dumontheil, 2014; Vendetti and Bunge, 2014).

Despite being associated with many abstract reasoning capabilities, there is yet to be an encompassing theory of rlPFC that explains its underlying computational contribution across domains. Hierarchical control theory has the most general account of rlPFC function, and suggests that it sits atop a rostrocaudal gradient of abstraction in the lateral cortex (Badre and D'Esposito, 2007, 2009). However, more recent work on control theory has challenged the notion that rlPFC is simply involved in abstract control processes, based on evidence that the rlPFC activates for a number of processes that are not easy to align with the idea that it is solely involved in abstract control (Badre and Nee, 2018). For example, it tends to be activated when people make the decision to explore new choice options rather than exploit options with the current highest expected value in reinforcement learning (Daw et al., 2006) and it has interesting temporal trajectories across sequential tasks that do not vary temporally in their control demands per se (Desrochers et al., 2015, 2019).

Recent research in category learning may shed light on how the rlPFC diverges from the other, more caudal, areas of the lateral PFC. Converging evidence across a number of studies suggests that the rlPFC instantiates an inferential process that is sensitive to the novelty and decisional uncertainty associated with a stimulus. Specifically, people will tend to engage rules when confronted with a stimulus that is both novel and difficult to categorize given the previously learned category representations. In support of this hypothesis, using an iterative rule-learning task, Paniukov and Davis (2018) found that the rlPFC is engaged early in learning and remains engaged as long as uncertainty remains about the correct category rule. Davis et al. (2017) showed that the rlPFC was more engaged during acquisition of relational category-learning rules than acquisition of feature-based category rules. Later, once the relational rules were well learned, the rlPFC was not more activated for relational rules across the

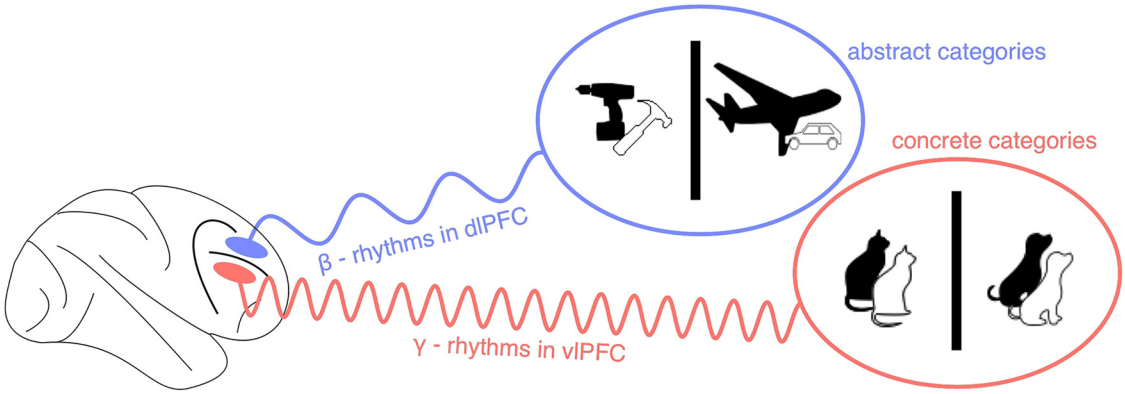

Figure 2. A schematic depiction of results from Wutz et al. (2018). Distinct neural circuits communicate through distinct channels when concept learning requires low levels of abstraction versus high levels of abstraction.

board, but instead was more activated to the extent that stimuli were novel examples of the previously learned relations. Finally, O'Bryan et al. (2018) showed that the rlPFC tracked the contribution of dissimilarity-based heuristics during generalization to novel ambiguous items. Such dissimilarity-based processes are thought to be based on higher-level inferential strategies. Together, these results suggest that the rlPFC integrates decisional information and stimulus novelty to determine when to use inferential processes in category learning and generalization.

Different prefrontal cortex dynamics for learning at different levels of abstraction

Many studies support the view that the lateral PFC plays a central role for concepts when they require high levels of abstraction (Badre and D'Esposito, 2009) and rule-based reasoning (Davis et al., 2017; O'Bryan et al., 2018). But lateral PFC has also been implicated in low-level, similarity-based categorization (Davis et al., 2017; O'Bryan et al., 2018) and representation of specific exemplars (Mack et al., 2013). Bowman and Zeithamova (2018, 2019) suggested that distinct processes and types of representations can be involved in concept learning, with their relative contribution depending on the coherence among category members. Lower coherence (lower similarity among category members) makes the formation of generalized representations based on similarity difficult and may require a higher level of abstraction. However, are high-level abstractions simply "more of the same", using the same mechanisms and networks as low-level abstractions? Or do low- and high-level abstractions involve distinct anatomical circuits and functional mechanisms in the PFC?

Important insights for resolving this question can come from neurophysiological recordings, which provide single-neuron resolution, but can also provide information about how groups of neurons communicate through analysis of their collective activity in local field potentials. To investigate how category abstraction is organized in prefrontal cortex, Wutz et al. (2018) recorded from multielectrode arrays in lateral PFC and trained monkeys in a dot-pattern category task (Posner and Keele, 1968) that provides a parametrically controlled and mathematically straightforward way to separate categories by level of abstraction. On each recording day, the monkeys learned to categorize dot patterns (i.e., exemplars), which were created by spatially distorting an underlying prototype pattern. Two new, pseudorandomly generated category prototypes were used per recording day, each of which was organized in a sequence of training blocks containing an increasing number of exemplars. By this, Wutz et al. (2018) made sure that the monkeys generalized over a large pool of exemplars (i.e., 64–256 per category) and eventually learned to extract the underlying category prototype. Critically, varying the degree of spatial distortion of the exemplars from the prototype allowed the authors to control the required level of abstraction for the category decision. Low-distortion exemplars look alike and can be categorized based on the similarity of their sensory features. High-distortion exemplars, however, can look very different from each other requiring greater abstraction.

Wutz et al. (2018) found category abstraction organized in different subregions in the PFC. The ventrolateral PFC (vlPFC) was more engaged for low-level abstractions, whereas the dorsolateral PFC (dlPFC) processed more high-level abstractions. Beyond this anatomical distinction, however, the combined analyses of spiking activity and local field potentials also suggested distinct functional mechanisms based on different temporal dynamics and frequency characteristics. In terms of frequency, there is growing evidence that gamma versus beta oscillations are involved in bottom-up versus top-down processing, respectively (Buschman and Miller, 2007; Jensen et al., 2007; Engel and Fries, 2010). Category signals for low-level abstractions in the vlPFC were found in gamma oscillations (60–160 Hz), evoked potentials, and spiking activity when the exemplars were shown and their sensory properties were processed. In contrast, the dlPFC showed stronger category signals for high-level abstractions in beta oscillations (10–35 Hz) throughout the memory delay epoch when the exemplars had to be kept in mind. Moreover, spiking activity in the dlPFC during the delay occurred at specific phases of its ongoing beta-band oscillations. This pattern suggested that neurons in the vlPFC were more driven by bottom-up inputs, whereas in the dlPFC neurons fired more in sync with its inherent, top-down dynamics during category processing.

The study demonstrated that distinct neural circuits (vlPFC vs dlPFC) communicate through distinct frequency channels (gamma vs beta) and at different times (sample vs delay epoch) when inferring regularities about the world on low versus high levels of abstraction (Fig. 2). By extension, these findings support the existence of two distinct functional mechanisms (bottom-up vs top-down) for category abstraction in the PFC. Gamma oscillations are typically linked to the feedforward flow of cortical activation (Bastos et al., 2015; Jensen et al., 2015) and the vlPFC receives direct inputs from inferior temporal cortex (O'Reilly, 2010), potentially continuing its functional properties into prefrontal cortex function. Therefore, category processing through the vlPFC-gamma network may be viewed as an object-

recognition/pattern-matching problem governed by bottom-up principles and subserving low-level abstraction. By contrast, beta oscillations support cortical feedback (Bastos et al., 2015; Jensen et al., 2015) and there is stronger connectivity between the dlPFC and parietal cortex (O'Reilly, 2010). Thus, the dlPFC-beta network may implement top-down, experience-based generalization and identify more abstract, conceptual relationships that transcend appearance beyond object recognition. Given that different regions of the lateral PFC contribute to category formation through distinct functional circuits because of their differential connectivity to posterior cortex, a critical goal for future research will be to delineate whether the lateral PFC is organized on the basis of representation type (rule vs similarity-based) or if more general control mechanisms define its topography (e.g., gating or branching; Badre and Nee, 2018).

In conclusion, the current review highlights representative samples from a recent boom of concept learning studies. By leveraging well developed computational models to interrogate neural mechanisms and representations, this work has implicated a broad network of brain regions including the hippocampus, PFC, and parietal cortices. Importantly, this work has significantly advanced our understanding of concept learning by characterizing the nature of the component mechanisms and their underlying neural machinery. The result is a converging neurocomputational account of concept learning that integrates brain systems involved in attention, memory, reasoning, cognitive control, and reward processing. Theoretically, this work brings resolution to the decades-long debate on the nature of category representations and emphasizes the need for comprehensive theories that bridge brain and behavior. Looking forward, understanding how these multiple brain systems interact throughout learning to support the flexible formation and use of concept knowledge is a key research aim. Notably, research in concept learning is well positioned to reach the holy grail of the computational model-based neuroscientific approach: not only are formal theories used to inform understanding of neural mechanisms, findings from the brain will undoubtedly motivate meaningful extensions to formal theories.

References

Aizenstein HJ, MacDonald AW, Stenger VA, Nebes RD, Larson JK, Ursu S, Carter CS (2000) Complementary category learning systems identified using event-related functional MRI. J Cogn Neurosci 12:977–987.

Aly M, Turk-Browne NB (2016) Attention promotes episodic encoding by stabilizing hippocampal representations. Proc Natl Acad Sci U S A 113:E420–E429.

Anderson MC (2003) Rethinking interference theory: executive control and the mechanisms of forgetting. J Mem Lang 49:415–445.

Ashby FG, Alfonso-Reese LA, Turken AU, Waldron EM (1998) A Neuropsychological Theory of Multiple Systems in Category Learning. Psychological Review 105:442–481.

Ashby FG, Maddox WT (2005) Human category learning. Annual review of psychology 56:149–178.

Ashby FG, Maddox WT (2011) Human category learning 2.0. Ann N Y Acad Sci 1224:147–161.

Badre D, D'Esposito M (2007) Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. J Cogn Neurosci 19:2082–2099.

Badre D, D'Esposito M (2009) Is the rostro-caudal axis of the frontal lobe hierarchical? Nat Rev Neurosci 10:659–669.

Badre D, Nee DE (2018) Frontal cortex and the hierarchical control of behavior. Trends Cogn Sci 22:170–188.

Badre D, Wagner AD (2005) Frontal lobe mechanisms that resolve proactive interference. Cereb Cortex 15:2003–2012.

Baldassano C, Hasson U, Norman KA (2018) Representation of real-world event schemas during narrative perception. J Neurosci 38:9689–9699.

Bastos AM, Vezoli J, Bosman CA, Schoffelen JM, Oostenveld R, Dowdall JR, De Weerd P, Kennedy H, Fries P (2015) Visual areas exert feedforward and feedback influences through distinct frequency channels. Neuron 85:390–401.

Berkman ET, Hutcherson CA, Livingston JL, Kahn LE, Inzlicht M (2017) Self-control as value-based choice. Curr Dir Psychol Sci 26:422–428.

Bowman CR, Zeithamova D (2018) Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. J Neurosci 38:2605–2614

Bowman CR, Zeithamova D (2019) Training typicality and set size effects on concept generalization and recognition. PsychXiv. Advance online publication. Retrieved June 1, 2019. doi:10.31234/osf.io/7g2q5.

Braunlich K, Seger CA (2016) Categorical evidence, confidence, and urgency during probabilistic categorization. Neuroimage 125:941–952.

Braunlich K, Liu Z, Seger CA (2017) Occipitotemporal category representations are sensitive to abstract category boundaries defined by generalization demands. J Neurosci 37:7631–7642.

Brod G, Lindenberger U, Shing YL (2017) Neural activation patterns during retrieval of schema-related memories: differences and commonalities between children and adults. Dev Sci 20:e12475.

Bruner JS, Goodnow JJ, Austin GA (1967) A study of thinking. New York: Wiley.

Bunge SA, Wendelken C, Badre D, Wagner AD (2005) Analogical reasoning and prefrontal cortex: evidence for separable retrieval and integration mechanisms. Cereb Cortex 15:239–249.

Burgess PW, Dumontheil I, Gilbert SJ (2007) The gateway hypothesis of rostral prefrontal cortex (area 10) function. Trends Cogn Sci 11:290–298.

Buschman TJ, Miller EK (2007) Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. Science 315:1860–1862.

Casale MB, Ashby FG (2008) A role for the perceptual representation memory system in category learning. Percept Psychophys 70:983–999.

Christoff K, Prabhakaran V, Dorfman J, Zhao Z, Kroger JK, KHolyoak KJ, Gabrieli JD (2001) Rostrolateral prefrontal cortex involvement in relational integration during reasoning. Neuroimage 14:1136–1149.

Clithero JA, Rangel A (2014) Informatic parcellation of the network involved in the computation of subjective value. Soc Cogn Affect Neurosci 9:1289–1302.

Davis T, Love BC, Preston AR (2012a) Striatal and hippocampal entropy and recognition signals in category learning: simultaneous processes revealed by model-based fMRI. J Exp Psychol Learn Mem Cogn 38:821–839.

Davis T, Love BC, Preston AR (2012b) Learning the exception to the rule: model-based fMRI reveals specialized representations for surprising category members. Cereb Cortex 22:260–273.

Davis T, Goldwater M, Giron J (2017) From concrete examples to abstract relations: the rostrolateral prefrontal cortex integrates novel examples into relational categories. Cereb Cortex 27:2652–2670.

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876–879.

Desrochers TM, Chatham CH, Badre D (2015) The necessity of rostrolateral prefrontal cortex for higher-level sequential behavior. Neuron 87:1357–1368.

Desrochers TM, Collins AG, Badre D (2019) Sequential control underlies robust ramping dynamics in the rostrolateral prefrontal cortex. J Neurosci 39:1471–1483.

Dumontheil I (2014) Development of abstract thinking during childhood and adolescence: the role of rostrolateral prefrontal cortex. Dev Cogn Neurosci 10:57–76.

Dusek JA, Eichenbaum H (1997) The hippocampus and memory for orderly stimulus relations. Proc Natl Acad Sci U S A 94:7109–7114.

Engel AK, Fries P (2010) Beta-band oscillations: signalling the status quo? Curr Opin Neurobiol 20:156–165.

Erickson MA, Kruschke JK (1998) Rules and exemplars in category learning. J Exp Psychol Gen 127:107–140.

Fenton AA, Lytton WW, Barry JM, Lenck-Santini PP, Zinyuk LE, Kubík S, Bures J, Poucet B, Muller RU, Olypher AV (2010) Attention-like modulation of hippocampus place cell discharge. J Neurosci 30:4613–4625.

Frank L, Bowman C, Zeithamova D (2019) Differential functional connectivity along the long axis of the hippocampus aligns with differential role in memory specificity and generalization. J Cogn Neurosci. Advance online publication. Retrieved August 9, 2019. doi:10.1162/jocn_a_01457.

Freedman DJ, Assad JA (2006) Experience-dependent representation of visual categories in parietal cortex. Nature 443:85–88.
Freedman DJ, Assad JA (2016) Neuronal mechanisms of visual categorization: an abstract view on decision making. Annu Rev Neurosci 39: 129–147.
Ghosh VE, Gilboa A (2014) What is a memory schema? A historical perspective on current neuroscience literature. Neuropsychologia 53:104–114.
Gilboa A, Marlatte H (2017) Neurobiology of schemas and schema-mediated memory. Trends Cogn Sci 21:618–631.
Gilboa A, Moscovitch M (2017) Ventromedial prefrontal cortex generates pre-stimulus theta coherence desynchronimzation: a schema instantiation hypothesis. Cortex 87:16–30.
Green AE, Fugelsang JA, Kraemer DJ, Shamosh NA, Dunbar KN (2006) Frontopolar cortex mediates abstract integration in analogy. Brain Res 1096:125–137.
Guise KG, Shapiro ML (2017) Medial prefrontal cortex reduces memory interference by modifying hippocampal encoding article medial prefrontal cortex reduces memory interference by modifying hippocampal encoding. Neuron 94:183–192.e8.
Hampshire A, Thompson R, Duncan J, Owen AM (2011) Lateral prefrontal cortex subregions make dissociable contributions during fluid reasoning. Cereb Cortex 21:1–10.
Homa D (1973) Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. J Exp Psychol 101:116–122.
Jensen O, Kaiser J, Lachaux JP (2007) Human gamma-frequency oscillations associated with attention and memory. Trends Neurosci 30:317–324.
Jensen O, Bonnefond M, Marshall TR, Tiesinga P (2015) Oscillatory mechanisms of feedforward and feedback visual processing. Trends Neurosci 38:192–194.
Jocham G, Furlong PM, Kröger IL, Kahn MC, Hunt LT, Behrens TE (2014) Dissociable contributions of ventromedial prefrontal and posterior parietal cortex to value-guided choice. Neuroimage 100:498–506.
Juslin P, Wennerholm P, Winman A (2001) High-level reasoning and base-rate use: do we need cue-competition to explain the inverse base-rate effect? J Exp Psychol Learn Mem Cogn 27:849–871.
Kable JW, Glimcher PW (2009) The neurobiology of decision: consensus and controversy. Neuron 63:733–745.
Knowlton BJ, Squire LR (1993) The learning of categories: parallel brain systems for item memory and category knowledge. Science 262:1747–1749.
Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis: connecting the branches of systems neuroscience. Front Syst Neurosci 1–28.
Kroger JK, Sabb FW, Fales CL, Bookheimer SY, Cohen MS, Holyoak KJ (2002) Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: a parametric study of relational complexity. Cereb Cortex 12: 477–485.
Kuhl BA, Chun MM (2014) Successful remembering elicits event-specific activity patterns in lateral parietal cortex. J Neurosci 34:8051–8060.
Kuhl BA, Rissman J, Chun MM, Wagner AD (2011) Fidelity of neural reactivation reveals competition between memories. Proc Natl Acad Sci U S A 108:5903–5908.
Kumaran D, McClelland JL (2012) Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. Psychol Rev 119:573–616.
Kumaran D, Summerfield JJ, Hassabis D, Maguire EA (2009) Tracking the emergence of conceptual knowledge during human decision making. Neuron 63:889–901.
Love BC, Gureckis TM (2007) Models in search of a brain. Cogn Affect Behav Neurosci 7:90–108.
Love BC, Medin DL, Gureckis TM (2004) SUSTAIN: a network model of category learning. Psychol Rev 111:309–332.
Mack ML, Preston AR, Love BC (2013)) decoding the brain's algorithm for categorization from its neural implementation. Curr Biol 23:2023–2027.
Mack ML, Love BC, Preston AR (2016) Dynamic updating of hippocampal object representations reflects new conceptual knowledge. Proc Natl Acad Sci U S A 113:13203–13208.
Mack ML, Love BC, Preston AR (2018) Building concepts one episode at a time: the hippocampus and concept formation. Neurosci Lett 680:31–38.
Mack ML, Preston AR, Love BC (2019) Ventromedial prefrontal cortex compression during concept learning. bioRxiv 178145.
Medin DL, Schaffer MM (1978) Context theory of classification learning. Psychol Rev 85:207–238.
Minda JP, Smith JD (2001) Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. J Exp Psychol Mem Cogn 27:775–799.
Morrison RG, Reber PJ, Bharani KL, Paller KA (2015) Dissociation of category-learning systems via brain potentials. Front Hum Neurosci 9:389.
Nomura EM, Maddox WT, Filoteo JV, Ing AD, Gitelman DR, Parrish TB, Mesulam MM, Reber PJ (2007) Neural correlates of rule-based and information-integration visual category learning. Cereb Cortex 17:37–43.
Nosofsky RM (1986) Attention, similarity, and the identification–categorization relationship. J Exp Psychol Gen 115:39–61.
Nosofsky RM, Palmeri TJ, McKinley SC (1994) Rule-plus-exception model of classification learning. Psychol Rev 101:53–79.
O'Bryan SR, Worthy DA, Livesey EJ, Davis T (2018) Model-based fMRI reveals dissimilarity processes underlying base rate neglect. eLife 7: e36395.
O'Reilly RC (2010) The what and how of prefrontal cortical organization. Trends Neurosci 33:355–361.
Palmeri TJ, Nosofsky RM (1995) Recognition memory for exceptions to the category rule. J Exp Psychol Learn Mem Cogn 21:548–568.
Paniukov D, Davis T (2018) The evaluative role of rostrolateral prefrontal cortex in rule-based category learning. Neuroimage 166:19–31.
Place R, Farovik A, Brockmann M, Eichenbaum H (2016) Bidirectional prefrontal-hippocampal interactions support context-guided memory. Nat Neurosci 19:992–994.
Posner MI, Keele SW (1968) On the genesis of abstract ideas. J exp Psychol 77:353–363.
Preston AR, Eichenbaum H (2013) Interplay of hippocampus and prefrontal cortex in memory. Current Biology 23:R764–R773.
Richter FR, Chanales AJH, Kuhl BA (2016) Predicting the integration of overlapping memories by decoding mnemonic processing states during learning. Neuroimage 124:323–335.
Romero K, Barense MD, Moscovitch M (2019) Coherence and congruency mediate medial temporal and medial prefrontal activity during event construction. Neuroimage 188:710–721.
Rosch E, Mervis CB (1975) Family resemblances: studies in the internal structure of categories. Cogn Psychol 7:573–605.
Ryan JD, D'Angelo MC, Kamino D, Ostreicher M, Moses SN, Rosenbaum RS (2016) Relational learning and transitive expression in aging and amnesia. Hippocampus 26:170–184.
Sarma A, Masse NY, Wang XJ, Freedman DJ (2016) Task-specific versus generalized mnemonic representations in parietal and prefrontal cortices. Nat Neurosci 19:143–149.
Schapiro AC, Turk-Browne NB, Botvinick MM, Norman KA (2017) Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. Philos Trans R Soc Lond B Biol Sci 372:20160049.
Schlichting ML, Preston AR (2015) Memory integration: neural mechanisms and implications for behavior. Curr Opin Behav Sci 111:1–18.
Schlichting ML, Preston AR (2016) Hippocampal-medial prefrontal circuit supports memory updating during learning and post-encoding rest. Neurobiol Learn Mem 134:91–106.
Schlichting ML, Zeithamova D, Preston AR (2014) CA1 subfield contributions to memory integration and inference. Hippocampus 24:1248–1260.
Schlichting ML, Mumford JA, Preston AR (2015) Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. Nat Commun 6:8151.
Seger CA, Miller EK (2010) Category learning in the brain. Annu Rev Neurosci 33:203–219.
Seger CA, Peterson EJ (2013) Categorization = decision making + generalization. Neurosci Biobehav Rev 37:1187–1200.
Seger CA, Braunlich K, Wehe HS, Liu Z (2015) Generalization in category learning: the roles of representational and decisional uncertainty. J Neurosci 35:8802–8812.
Sestieri C, Shulman GL, Corbetta M (2017) The contribution of the human posterior parietal cortex to episodic memory. Nat Rev Neurosci 18:183–192.

Smith EE, Sloman SA (1994) Similarity vs. rule-based categorization. Mem Cogn 22:377–386.

Spalding KN, Jones SH, Duff MC, Tranel D, Warren DE (2015) Investigating the neural correlates of schemas: ventromedial prefrontal cortex is necessary for normal schematic influence on memory. J Neurosci 35: 15746–15751

Spalding KN, Schlichting ML, Zeithamova D, Preston AR, Tranel D, Duff MC, Warren DE (2018) Ventromedial prefrontal cortex is necessary for normal associative inference and memory integration. J Neurosci 38: 3767–3775.

Squire LR, Knowlton BJ (1995) Learning about categories in the absence of memory. Proc Natl Acad Sci U S A 92:12470–12474.

Summerfield C, Tsetsos K (2012) Building bridges between perceptual and economic decision-making: neural and computational mechanisms. Front Neurosci 6:70.

Tse D, Langston RF, Kakeyama M, Bethus I, Spooner PA, Wood ER, Witter MP, Morris RG (2007) Schemas and memory consolidation. Science 316:76–82.

Tse D, Takeuchi T, Kakeyama M, Kajii Y, Okuno H, Tohyama C, Bito H, Morris RG (2011) Schema-dependent gene activation and memory encoding in neocortex. Science 333:891–895.

Underwood BJ (1949) Proactive inhibition as a function of time and degree of prior learning. J Exp Psychol 39:24–34.

van Kesteren MTR, Krabbendam L, Meeter M (2018) Integrating educational knowledge: reactivation of prior knowledge during educational learning enhances memory integration. NPJ Sci Learn 3:11.

van Kesteren MTR, Ruiter DJ, Fernández G, Henson RN (2012) How schema and novelty augment memory formation. Trends Neurosci 35:211–219.

van Kesteren MTR, Rijpkema M, Ruiter DJ, Fernández G (2013a) Consolidation differentially modulates schema effects on memory for items and associations. PLoS One 8:e56155.

van Kesteren MTR, Beul SF, Takashima A, Henson RN, Ruiter DJ, Fernández G (2013b) Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: from congruent to incongruent. Neuropsychologia 51:2352–2359.

van Kesteren MTR, Rignanese R, Gianferrara PG, Krabbendam L, Meeter M (2019) Integrating memories: Congruency and reactivation aid memory integration through reinstatement of prior knowledge. bioRxiv 716076.

Vendetti MS, Bunge SA (2014) Evolutionary and developmental changes in the lateral frontoparietal network: a little goes a long way for higher-level cognition. Neuron 84:906–917.

Watson CE, Chatterjee A (2012) A bilateral frontoparietal network underlies visuospatial analogical reasoning. Neuroimage 59:2831–2838.

Wheeler ME, Woo SG, Ansel T, Tremel JJ, Collier AL, Velanova K, Ploran EJ, Yang T (2015) The strength of gradually accruing probabilistic evidence modulates brain activity during a categorical decision. J Cogn Neurosci 27:705–719.

Wutz A, Loonis R, Roy JE, Donoghue JA, Miller EK (2018) Different levels of category abstraction by different dynamics in different prefrontal areas. Neuron 97:716–726.e8.

Zeithamova D, Maddox WT, Schnyer DM (2008) Dissociable prototype learning systems: evidence from the brain and behavior. J Neurosci 28:13194–13201.

Zeithamova D, Dominick AL, Preston AR (2012) Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. Neuron 75:168–179.

Zeithamova D, Preston AR (2017) Temporal proximity promotes integration of overlapping events. J Cogn Neurosci 29:1311–1323