Schema, Inference, and Memory

Nicole L. Varga, Ph.D.[1], Neal W Morton, Ph.D.[1], & Alison R. Preston, Ph.D.[1,2,3]

Affiliations

[1]Center for Learning & Memory, The University of Texas at Austin, United States, [2]Department of Psychology, The University of Texas at Austin, United States, [3]Department of Neuroscience, The University of Texas at Austin, United States

Correspondence

Nicole L. Varga (nvarga@austin.utexas.edu) & Alison R. Preston (apreston@utexas.edu)

To appear in: Oxford Handbook of Human Memory
Oxford University Press
Michael J. Kahana & Anthony D. Wagner, editors
(Table of contents: https://memory.psych.upenn.edu/Oxford_Handbook_of_Human_Memory)

Abstract

The question of how knowledge structures, or schemas, are formed and how they influence memory and inference has posed long-standing challenges for cognitive scientists. Recent neuroscientific advances have improved our ability to quantify schemas as they are formed and during their expression in novel situations, thus improving our mechanistic understanding of their role in cognition. Here, we review recent evidence indicating that bidirectional interactions between the hippocampus and medial prefrontal cortex (mPFC) mediate schema formation, promoting creation of knowledge representations that integrate commonalities and emphasize goal-related differences among related events. We discuss how bidirectional hippocampus—mPFC interactions control memory reactivation by preferentially retrieving information that overlaps with current experience. Moreover, we review evidence for hippocampal binding processes that link current experience with reactivated memories, thereby supporting schema formation. We further focus on new data revealing that mPFC may guide formation of hippocampal schemas, by biasing encoding toward goal-relevant information while compressing irrelevant features that differ across events. Critically, by representing the commonalities and differences among previously and newly acquired information, schemas extend beyond direct experience to support flexible behaviors, such as inferential reasoning. Optimal behavior in familiar and novel contexts may thus be rooted in the dynamic interplay between neural structures that support schema formation and expression. As such, this chapter considers current perspectives on schemas in the context of emerging neural evidence, providing new insight into the relationship between memory and reasoning, as well as the implications of recent neuroscientific work on schema theory and cognition more broadly.

Keywords: medial prefrontal cortex, hippocampus, integration, reactivation, compression, latent causes, hierarchical knowledge, reasoning

1. Introduction

Schemas are knowledge frameworks that capture the structure of the environment and allow one to predict the correct actions to take in both familiar and novel contexts (Bartlett, 1932; Piaget, 1954). Despite ongoing debate regarding the precise nature of schemas, current perspectives generally agree that schemas extract goal-relevant features that are common across individual episodes (Ghosh & Gilboa, 2014; Preston & Eichenbaum, 2013; van Kesteren et al., 2012; Wang & Morris, 2010). We propose that schemas consist of context-specific associative structures that are abstracted across multiple events. Once formed, schemas further support inference and generalization. For example, across multiple visits to restaurants, formation of a restaurant schema would capture the common temporal structure across visits that reliably supports one's goal of acquiring food. Upon entering a new restaurant, retrieval of one's restaurant schema may allow one to infer that the presence of a host stand means that one should wait for the host to return before seating oneself. Furthermore, one could use schema knowledge to predict the event sequences that should follow next, including being handed a menu at one's table, selecting a meal from the menu, placing an order with the server, and being served food. Schema formation thus promotes optimal behavior by minimizing uncertainty about how to act in new settings.

While all theoretical models of schemas acknowledge the importance of commonalities to prediction and generalization (Ghosh & Gilboa, 2014), representation of differences is also essential to determining optimal behavior. Consider sit-down and fast food restaurants, which contain similar goal-relevant actions but different underlying temporal structures. Unlike the sit-down schema which dictates that one should wait to be seated, in a fast food restaurant, one should order food and wait at the counter for it before sitting at a table. Hence, if an individual deployed the set of actions determined by their sit-down restaurant schema in the fast food context, they would wait to be seated without ever attaining their goal of food. We propose that schemas are hierarchically organized (Eichenbaum, 2017), with contextual features differentiating the precise sequence of actions to take in any given environment.

Schemas therefore preferentially represent goal-relevant commonalities and differences that are critical to making predictions about behavior. Features of events that are neither common across episodes nor predictive are not represented in schemas. For instance, a one-time event in which the waiter drops your plate as he approaches your table would not be included in one’s restaurant schema. Although the dropped plate has consequences for one’s immediate goal of eating food, the specific event is not likely to predict what one should expect to happen during a future trip to a restaurant, and thus would not be represented in the restaurant schema. Schemas thus serve to reduce the vast amount of information in the environment, efficiently representing only those environmental features that support accurate generalization during future events.

Once formed, schemas influence how new information is learned and represented. They evolve over time to incorporate new, predictive knowledge and also change when previously learned information is no longer accurate. When new experiences are similar to schema predictions, learning may be speeded as new content can be readily assimilated into existing schemas. For example, when visiting an expensive restaurant for the first time, one may be handed hot towels before and after the meal to cleanse one’s hands. These new events may be incorporated into the temporal structure represented within a restaurant schema so that one might expect to receive hot towels at the beginning and end of a meal in the future. Schema assimilation thus allows individuals to encode and organize new event information in the context of existing knowledge.

Whereas assimilation updates schemas to incorporate new schema-consistent content, accommodation involves restructuring of schemas to account for new experiences that deviate from what one expects based on current schemas. To promote optimal behavior in a host of situations, schemas must be refined when current knowledge structures no longer inform how to behave optimally in new settings. Consider the introduction of fast casual restaurants, in which the sequence of actions necessary to receive a meal is neither predicted by one's previous experiences at sit-down restaurants nor by experiences at fast food restaurants. The structure of expectations for behavior therefore differs fundamentally among the restaurant types, and thus assimilation into the existing schema is not possible. If an individual deployed the set of actions determined by their sit-down predictions in a fast-casual restaurant, they would not receive food because they failed to order it at a counter. To better navigate this new scenario, the existing schema must be accommodated (i.e., changed). That is, new temporal and contextual features must be represented to support differentiation of the action predictions in each type of restaurant. Differentiation thus enables formation of more complex hierarchies of schematic knowledge to accommodate prediction in the widest variety of circumstances.

In this chapter, we review recent neuroscientific and computational work on schemas. We first review the historical roots of schema theory, highlighting universal psychological principles that continue to influence contemporary research on schemas today. We further describe the longstanding challenges in using psychological terms to define the nature of schemas and set forth a neurocomputational framework that aims to define schemas in more mechanistic terms (see **Figure 1.1**). In particular, we review recent neuroscience studies that use representational analysis strategies to quantify the organization of schemas during their initial formation and later expression in novel situations. We further review evidence indicating that the hippocampus represents the predictive structure of individual experiences (e.g., temporal structure of a restaurant experience) and binds current experience with reactivated memories (e.g., extracting commonalities across individual restaurant experiences), thereby supporting schema formation. We further discuss how hippocampal separation mechanisms allow for differentiation within schemas, allowing for hierarchical representation of context-specific predictions (e.g., sit-down versus fast food restaurants). Finally, we review recent data revealing that medial prefrontal cortex (mPFC) interactions with hippocampus play a key role in controlling schema formation and expression, by biasing encoding toward goal-relevant features and reactivating relevant knowledge during expression. Through delineating the basic building blocks that give rise to complex schematic structures, we provide a mechanistic framework for how schemas support optimal behavior in new situations including memory, reasoning, and decision making.

2. Contemporary schema theory is grounded in historical roots

A key principle of both historical and contemporary schema theories is that schemas exert a profound impact on new learning and memory. Original work on schemas focused on how individuals remembered structured narratives, with the aim of understanding how prior knowledge influenced formation and retrieval of new memories. For instance, Bartlett (1932) showed that repeated attempts to recall a complex Native American folktale depended on individuals' prior experience of Western folktales. When retelling the Native American folktale from memory, participants frequently omitted events from the story that were atypical relative to the common structure of Western folktales, such as the depiction of a spirit leaving a character's body. This finding suggests that when retrieving the Native American folktale, participants reconstructed the original episode through the lens of their existing knowledge of Western folktales. Events from the Native American story that were not predicted from their Western folktale schema were less likely to be encoded into or retrieved from memory (see Alba & Hasher, 1983 for a comprehensive review of the effects of schemas on memory encoding and retrieval). This seminal work demonstrating how expectations based on prior knowledge guide how we interpret and remember new events continues to inspire modern theories of schema representation and expression today (van Kesteren et al., 2013; van Kesteren et al., 2014).

A second principle of schemas is that they guide prediction and inference, which was originally supported by early work on spatial navigation in rodents. Tolman (1948) proposed that rodents construct a "cognitive map" of the environment, which can be thought of as a spatial schema that represents individual objects and both their locations in the world and locations relative to one another. Those spatial relationships among objects may be learned directly, by traversing from one object to another in a single trip along a path, or through inference, by linking information acquired through multiple trips through the environment. Formation of such spatial schemas facilitates behavior during subsequent experiences within the environment. For example, Tolman showed that rodents take less time and make fewer errors when navigating to a reward on a previously traversed route through the environment, suggesting that they are able to use the learned spatial schema to predict the series of locations that lead to the goal location. Furthermore, by associating individually-traversed maze trajectories to form a comprehensive schema of the environment, rodents are able to make inferences about new navigational routes. For example, when previously learned paths were blocked in the environment, rodents were able to quickly find an alternative route that represented the optimal path to a goal location, even if that trajectory had not been directly experienced previously. Schema formation therefore promotes optimal behavior by promoting inference and minimizing uncertainty about how to act in new settings.

A third principle of schemas is that they are highly adaptable; schemas evolve continually, changing to incorporate new information and to more efficiently represent the structure of the world. Piaget's work in cognitive development provided an influential framework for understanding how schemas are updated or altered to incorporate new experiences (Piaget, 1972). Specifically, Piaget introduced the concepts of *assimilation* and *accommodation* as two processes through which schema knowledge may change. When new events are consistent with predictions derived from existing schemas, they are assimilated into that knowledge, providing additional evidence about predictive commonalities across events. However, when new events differ from schema predictions, accommodation must occur. During accommodation, either the knowledge contained in existing schemas is altered to make new predictions, or a new schema is created. While the psychological constructs of assimilation and accommodation remain influential today, Piaget's descriptions of the mechanisms whereby we extract and internally represent commonalities and differences across events were far from precise. In subsequent sections, we present the challenges of testing schema theory and review recent neuroscientific work that has offered greater clarity on the mechanisms of schema formation, expression, and updating.

3. Longstanding challenges of schema theory

While the principles of early schema theories provide a framework for understanding the cognitive processes that schemas support, such as memory and inference, there has long been substantial controversy about the precise, operational definition of a schema. One reason for this ongoing challenge is that both seminal and contemporary work on schemas tend to interpret memory and inference behavior through the lens of the unobservable psychological construct of a schema in the absence of direct measurement. For instance, Bartlett presumed that an activated schema influenced the reconstructive memory behavior he observed, and thus inferred its definition as “an active organization of past reactions, or of past experiences, which must always be supposed to be operating” (Bartlett, 1932). The continued tendency to define an internal psychological structure based on observable behavior poses substantial difficulty to moving beyond imprecise theoretical constructs. Even Bartlett admitted that he disliked the term schema, arguing that “it is at once too definite and too sketchy.” Indeed, Bartlett’s hesitation in adopting the word schema stemmed from the observation that it was “already widely used in controversial psychological writing to refer generally to any rather vaguely outlined theory.” Defining schemas in terms of the schema-like behavioral effects reported across the 90 years since the term’s introduction therefore necessitates a vague definition to encompass a wide variety of findings. Such imprecise definitions raise challenges for moving schema theory beyond a description of behavioral effects on memory and inference, particularly when considering the possible neural instantiation of schemas.

To extend schema theory beyond descriptions of behavior, some theoretical models have attempted to provide more precise accounts of how schemas are instantiated. At one extreme, a class of models have proposed that a schema may be abstracted from the summed content of one’s corpus of specific memory traces at the time of expression (Hintzman, 1986; Restle, 1961). These models assume that schemas are not stored knowledge, but rather simply computed on the fly from individual memory traces. Proponents of such retrieval-based accounts of schemas highlight the advantage of these models in avoiding the “vexing question of how schemas are learned” (Hintzman, 1986). In contrast, other prominent schema theories argue that these mental structures are represented in memory, and that new information is continuously integrated into existing schemas (Alba & Hasher, 1983). However, if this is the case, such a theory must also explain *how* new experience is incorporated into existing schemas. This chapter thus considers schema theory in the context of emerging neuroscientific work which has improved our ability to quantify schemas, offering new insight into the “vexing question” of how they are formed. By directly isolating the neural mechanisms that support schema formation and expression under controlled conditions, the approach has furthered our mechanistic understanding of schema theory and the role of schemas in cognition.

4. Neural mechanisms that support schema formation

In this section, we provide a neuroscientific and computational framework for how individuals come to represent the predictive structure of the world, a question which has posed long-standing challenges for schema theory. We build on the theoretical argument that schemas are a highly structured form of knowledge and directly reflect the systematic structure of our environment, owing to the fact that “*the perceived world is not an unstructured total set of equiprobably co-occurring attributes … [T]he material objects of the world possess high correlational structure*” (Rosch et al., 1976). Recent neuroscientific advances have allowed us to isolate mechanisms that support the core hypothesized principles of schemas, in particular by allowing us to measure representational knowledge structures as they are formed, expressed, and updated.

We focus on new findings from electrophysiological and human neuroimaging studies that use representational analysis approaches to quantify how the correlational structure of the world is reflected in neural representations. In addition, we review perspectives derived from computational models, which have provided powerful mathematical frameworks for explaining how incoming information from the world is reduced into schema representations that emphasize goal-relevant features of the environment. We further propose a mechanistic framework for how the brain expresses and updates schemas in new contexts to account for new environmental relationships. In doing so, we offer a mechanistic update to Piaget's concepts of assimilation and accommodation, demonstrating how specific computational and neural functions support these proposed processes.

4.1. Hippocampal binding supports context-specific associative predictions

At a basic level, schemas require formation of associative memories that link items and contexts to specific behaviors and outcomes. The auto-associative properties of the hippocampal network are thought to be critical for such binding (Manns & Eichenbaum, 2006; Marr, 1971; McClelland et al., 1995). For instance, recent work in statistical learning demonstrates that hippocampal representations become more similar for items that reliably occur one after another in a continuous sequence of events (Schapiro et al., 2012), reflecting how they are bound together in memory. Similarly, complimentary work in spatial navigation indicates that hippocampal representations become more similar for objects judged as more spatially proximal in a recently learned virtual environment (Deuker et al., 2016). Such representational binding is thought to support predictive reinstatement during memory retrieval to guide efficient decision making. When cued with a previously experienced stimulus, hippocampal pattern completion mechanisms result in reinstatement of the neural representation of bound stimuli in anticipation of the stimulus that should appear next in either time or space (Johnson & Redish, 2007; Karlsson & Frank, 2009; Redish, 2016; Stachenfeld et al., 2017).

Such anticipatory predictions are often context-specific—another key facet for schema representation. For example, a recent study used a context-specific prediction paradigm, in which the same cue stimulus (item A) when followed by different actions (a left or right button press) led to different outcomes (presentation of item B or presentation of item C). Hippocampal representations differentiated the sequences that predicted different stimulus outcomes, allowing for accurate reinstatement of an anticipated stimulus outcome after a cue item-action combination (Hindy et al., 2016). Furthermore, when associative relationships are less reliable, hippocampal representations become dissimilar for items that co-occur some but not all of the time (Schapiro et al., 2012), suggesting the mechanism through which content-specific differentiation might occur. While these findings concentrate on representation of directly observed associations, the basic binding and reinstatement mechanisms reviewed here serve as a foundation for forming more complex associative schemas that extract structure beyond direct experience.

4.2 Hippocampal associative mechanisms enable inference about commonalities among events

Extracting associative structure across multiple events—an inherent property of schemas—requires both reactivating memories for events that overlap with new experiences and binding new information with reactivated memories. Lesion work in both rodents (Bunsey &

Eichenbaum, 1996; Dusek & Eichenbaum, 1997) and humans (Pajkert et al., 2017) shows that the hippocampus is necessary for extracting commonalities across distinct events. These studies typically have employed associative or transitive inference paradigms to assess the impact of hippocampal lesions on cross-event associative knowledge acquisition. In one such study (Dusek & Eichenbaum, 1997), rodents learned a set of odor paired associates through trial-and-error reinforcement, whereby they were exposed to two items and were tasked with choosing which one led to reward (e.g., A > B). Across five different odors, there was a hierarchical relationship of reward outcomes, for which item A was always rewarded in the presence of B, B was always rewarded in the presence of C, etc. (i.e., A > B, B > C, C > D, D > E). As such, rodents could learn the directly experienced associations (e.g., A should be selected over B) as well as the full hierarchy of reward outcomes that would support inference about novel pair combinations during a critical knowledge test (e.g., B should be selected over D). Damage to the hippocampus resulted in preserved memory for the individual associations (e.g., B > C, C > D) but impaired performance on inferences that required knowledge of the full hierarchy (e.g., B > D). These findings indicate that hippocampal associative mechanisms are critical for extracting hierarchical structure across individual experiences.

Recent neuroimaging studies have provided additional mechanistic insight into how knowledge extraction across episodes occurs. A first step in forming knowledge schemas is reactivating previously experienced events that share features with current experiences, to allow for the extraction of commonalities (Morton et al., 2017; Schlichting & Preston, 2015). When a new experience overlaps with a prior episode, hippocampal pattern completion mechanisms trigger reactivation of the previous memory in both humans (Gershman, Schapiro, et al., 2013; Kuhl et al., 2012; Zeithamova et al., 2012) and animals (Ji & Wilson, 2007; Karlsson & Frank, 2009). Such reactivation promotes inference about unobserved relationships similar to those indexed in the rodent lesion work reviewed above (Shohamy & Wagner, 2008; Wimmer & Shohamy, 2012; Zeithamova et al., 2012). For instance, in one study (Zeithamova et al., 2012), participants learned overlapping pairs of associations (e.g., AB: chair & zucchini and BC: zucchini & blender) and were later asked to make judgements about the relationships between the two items that shared a common associate (i.e., AC inference: chair & blender). Those participants who reactivated related memory elements during overlapping event encoding (e.g., reactivating A when studying BC) were superior at making novel inferences. These findings suggest that reactivating related memories during learning promotes binding of the past with the present, leading to formation of integrated memory networks that represent associations beyond direct experience. Such knowledge extraction through integration may support prediction and inference in novel situations and likely represents the first step in schema formation (Preston & Eichenbaum, 2013).

More direct evidence for cross-episode hippocampal integration comes from recent human neuroimaging studies using multivariate analysis approaches to assess how representation of elements of overlapping memories shifts to reflect representation of commonalities (Collin et al., 2015; Mack et al., 2016; Schlichting et al., 2015). Two such studies showed that hippocampus represented indirectly related elements (i.e., A and C items) from overlapping events (AB and BC pairs) as more similar to one another after learning (Schlichting et al., 2015), with hippocampal integration increasing over time as a function of consolidation (Tompary & Davachi, 2017). Such representation of unobserved relationships within hippocampus has further been linked to increased ability to infer connections among individual events (Collin et al., 2015).

Within the hippocampus, the CA_1 subfield may play a particular role in formation of integrated memory networks that represent commonalities across experiences given its anatomical properties and patterns of connectivity. CA_1 receives input about reactivated memories from CA_3 and converging input about current sensory experience from entorhinal cortex (Suh et al., 2011; van Strien et al., 2009; Witter, 2011). As such, CA_1 simultaneously processes newly encoded and reactivated memory representations (Larkin et al., 2014; Lisman & Grace, 2005). The CA_1 region is thus well-situated to integrate incoming experience with previously acquired memories, promoting creation of knowledge representations that extract commonalities among separate events. In support of this idea, simulations generated from a computational model of the hippocampal circuit show that the CA_1 subfield uniquely represented temporal regularities and associative commonalities derived across multiple events (Schapiro et al., 2017). Reactivation of prior event details during new event encoding within human CA_1 has further been linked with a superior ability to infer connections among related events (Schlichting et al., 2014).

Perhaps the most direct evidence for the common representation of events within CA_1 comes from a contextual fear conditioning study in mice, in which the activity of multiple CA_1 neurons was imaged simultaneously (Cai et al., 2016). In this work, animals were exposed to an initial spatial context (context A) and a second context (context B) after several hours. Even though these events were separated by a long interval, the CA_1 ensembles active during these two events were highly overlapping. Moreover, when returned to the second context (B) two days later and given a foot shock, rodents transferred the learned fear response from the initial context (A) that shared an overlapping hippocampal representation, demonstrating how memory integration supports generalization. Importantly, when mice failed to form overlapping representations of the two spatial contexts within CA_1, generalization of fear was not observed.

Collectively, these results from both human and animal studies indicate that overlapping representations in hippocampus, and CA_1 in particular, allow for the representation of commonalities that go beyond direct experience to support both inference and generalization. While these studies inform the mechanisms that support formation of commonalities across a limited set of individual associations, the same processes are likely to be essential to the formation of more complex schematic relationships (Mack et al., 2018). As evidence for the hippocampus' ability to represent associative relationships among numerous stimuli, several recent studies indicate that information about the temporal, spatial, and conceptual distances between multiple objects can be read out from hippocampal activity patterns (Garvert et al., 2017; Mack et al., 2016; Schapiro et al., 2016).

4.3 Hippocampal differentiation supports hierarchical representation

Hierarchical representation may be an essential part of how schemas are organized. Representing commonalities across events allows for the extraction of general principles that can guide predictive behavior, but it is important that such predictions be context- and goal-specific. Consider again the differences between fast food and sit-down restaurants, which have similar goal-relevant actions but must be enacted in a different temporal order to successfully obtain one's meal. Within one's restaurant schema, it will therefore be necessary to differentiate among types of restaurants to accurately predict the relationships between items, actions, and outcomes. Recent electrophysiological data from rodents indicates that hippocampus forms hierarchical representations that simultaneously represent the commonalities among and differences between events to support such context-specific behavior (McKenzie et al., 2014).

In a context-guided object association task, rodents were presented with two sets of objects (AB and CD). On each trial, the objects from one of the sets were presented simultaneously, and the animals had to select the object that was associated with reward. Importantly, the association between an object and reward in this task was context dependent (**Figure 1.2A**). For instance, when objects A and B were presented in one context, A was always associated with reward. However, when the same objects were presented in a second, perceptually distinct context, the reward contingencies changed, with object B being associated with reward in the second context. The context-specific reward associations were shared across the two object sets, such that objects A and C were always rewarded in the same context, with B and D being rewarded in the opposing context (**Figure 1.2B**). Across trials, the positions of the two objects (i.e., whether they appeared on the left or right side of the context) further varied. This task design thus allowed the researchers to quantify how population activity within the hippocampus varied as a function of item, context, reward valence, and spatial position, enabling assessment of whether representation of the different task features were hierarchically organized (McKenzie et al., 2014).

The results indicated that hippocampal population activity was anticorrelated when rodents approached the same item in different contexts (**Figure 1.2C**). In other words, the same object elicited different patterns of hippocampal response when presented in the different contexts, reflecting the fact that different outcomes were associated with the same object in the two contexts. The hippocampus simultaneously represented similarities among events that occurred within the same context. Repeated presentation of the same object within the same context evoked the most similar responses within hippocampus, followed by objects associated with the same reward outcome (e.g., A and C), and objects in the same location (regardless of their identity or reward outcome). This pattern of response indicates hierarchical representation of the task properties in hippocampus, with context information being differentiated at the highest level of representation and commonalities by position, valence, and object being represented at successive levels (**Figure 1.2C**). As such, this observation stands as one of the clearest demonstrations of a representational schema within hippocampus to date. Importantly, the context specific associations represented in this hierarchy speeded learning of new object-context reward associations (McKenzie et al., 2014), demonstrating how schemas can facilitate acquisition of new content.

4.4 Latent cause models and schemas

In the examples above, relatively simple rules dictate how different actions in different contexts result in different outcomes. In the real world, however, the relationship between behavior and outcomes is rarely explained via simple rules. Schemas must therefore represent a number of multidimensional features that predict which behaviors are most adaptive in any given context. Latent cause models provide a computational means to represent multidimensional predictions, and thus may capture an essential representational requirement of schemas. Latent cause models propose that to resolve uncertainty about the structure of the world and exploit that structure in new situations, one must infer and represent the hidden causes of direct observations (Gershman et al., 2017). For example, to retrieve the correct schema when entering a restaurant, one may infer what type of restaurant it is (e.g., formal or casual) based on visual cues (e.g., seeing a patron wearing a suit). Latent causes may be thought of as an index to a multidimensional feature space that codes associative relationships among event elements, actions, and outcomes. Latent causes therefore tap into structured associative knowledge to acquire beliefs about the combination of associations that consistently lead to an observed outcome. Through acquiring a belief about the latent causes of observed outcomes, one may later use the belief to guide prediction about one's current state, and hence, generate predictions for how to behave optimally (see **Box 1.1**).

We next take a simple example from fear conditioning as a useful way of demonstrating how latent causes mediate representation of associative predictions. In Pavlovian reinforcement learning, a rodent is presented with a conditioned stimulus (e.g., a tone) followed by an unconditioned stimulus (e.g., a shock) that naturally elicits an unconditioned response (e.g., freezing). Following repeated exposures to the tone-shock pairing, the tone reliably elicits the freezing response even in the absence of a shock, indicating that the animal has learned the relationship among the stimulus and the outcome (Rescorla, 1988). The mechanistic explanation for fear conditioning offered within the latent cause model is that repeated tone-shock experiences are assigned to the same latent cause—a "fear acquisition" state (Gershman et al., 2017).

When a previously inferred latent cause is thought to be active again, such as during repeated exposures to the tone, it promotes retrieval of the memory associated with that latent cause (i.e., the collection of experiences with the tone-shock pairing). These latent structures are subsequently updated in accordance with how well the inferred latent structure accounts for current experience, resulting in strengthening of commonalities and weakening of idiosyncratic elements between overlapping experiences that are not predictive. In this case, the associative weights between the tone and shock underlying the fear acquisition state will strengthen over repeated experiences, promoting efficient retrieval and prediction of a shock when subsequently presented with a tone alone, thereby eliciting freezing.

As discussed above, schemas not only represent commonalities among events, but also differences. In line with this idea, the latent cause model provides a plausible explanation for how associative knowledge structures simultaneously represent key distinctions between highly overlapping experiences. Fear conditioning again provides a useful example for how differentiation guides formation of latent causes. During fear extinction, the animal's freezing response gradually diminishes when the tone is repeatedly presented alone, without a corresponding shock. Yet findings indicate that extinction of fear is temporary and recovers with the passage of time (Bouton, 2004). Instead of overwriting the original fear memory, the latent cause model suggests that the animal infers the presence of a new "extinction" state (Gershman et al., 2010).

During extinction, the presentation of the tone will activate a belief (based on the latent cause) that the shock should follow. When the predicted shock is not administered, it is beneficial to generate a new causal structure to adequately account for the unexpected lack of shock (Gershman et al., 2010; Gershman, Jones, et al., 2013). The result is two differentiable states that assign different latent causes to the situations; one in which the tone is followed by a shock versus one in which it is not. In other words, because latent structures are updated in accordance with how well the inferred state accounts for current experience, new latent causes will be inferred when predictions made from existing latent causes are violated. Recent empirical evidence shows that separate "fear" and "extinction" states are represented in different hippocampal ensembles during extinction (Lacagnina et al., 2019). Such empirical data further demonstrate how differentiation is a key facet through which we organize associative predictions and make inferences about expectations from that knowledge.

Latent cause models can scale up from relatively simple associative predictions, such as in the fear conditioning example above, to account for more complex associative spaces for which multidimensional task features need to be represented to accurately predict outcomes. For example, in one recent study (Chan et al., 2016), individuals were placed in a hypothetical zoo which had four different sectors (**Figure 1.2D**). The sectors of the zoo were differentiated by the probability of seeing different animals in each region. For instance, in the blue sector, participants were most likely to see elephants and giraffes and less likely to see lions and zebras, whereas, in the pink sector, the probability of seeing any one of those same animals was similar. Participants first learned about the individual associations between the zoo sectors and each animal, forming a likelihood distribution of the probability of seeing any one animal in a particular sector of the zoo (**Figure 1.2E**). These probability distributions, which reflect latent beliefs about the structure of the zoo, then allowed individuals to make inferences about their location in the zoo based solely on which animals they observed in a context (**Figure 1.2D, bottom panel**).

4.5 Medial prefrontal cortex interacts with hippocampus to extract latent causes

If hippocampal binding mechanisms support formation of structured associative knowledge that represents commonalities and differences across individual events, mPFC may tap into that associative knowledge to index beliefs about the combination of associations that consistently lead to a desired outcome (i.e., latent causes). Medial PFC has direct anatomical connections to the hippocampus, receiving inputs primarily from the anterior portion of CA_1

(Barbas & Blatt, 1995; Cavada et al., 2000). Medial PFC also has extensive connections with a diverse set of sensory, limbic, and subcortical structures (Cavada et al., 2000). These anatomical properties of mPFC may make it especially well suited at indexing how multidimensional features of the environment relate to one's actions and goals. Accordingly, computations and representations supported by both hippocampus and mPFC may be essential to schemas, with their interactions determining how schemas are formed and later accessed (Preston & Eichenbaum, 2013; Schlichting & Preston, 2015; Wikenheiser & Schoenbaum, 2016).

Consistent with this idea, mPFC—hippocampal interactions in both rodents (Jadhav et al., 2016; Yu et al., 2018) and humans (Liu et al., 2017; Schlichting & Preston, 2016; Zeithamova et al., 2012) are enhanced when new events overlap with existing knowledge, with their connectivity predicting individuals' ability to infer relationships among discrete events (Schlichting & Preston, 2016; Zeithamova et al., 2012). Moreover, while lesions to mPFC do not impair direct learning of simple associations, they do impair inference about across-event relationships (Koscik & Tranel, 2012; Spalding et al., 2018) and individuals' ability to learn from observed outcomes (Kumaran et al., 2015). These lesion findings suggest that mPFC may draw upon associative input from hippocampus to extract information about associative features that lead to specific outcomes. In context-dependent learning tasks, information about the multidimensional features that predict different outcomes is reflected in the population activity of mPFC neurons (Farovik et al., 2015; Wikenheiser et al., 2017). However, when hippocampal inputs to mPFC are temporally inactivated, predictive reinstatement of expected outcomes in mPFC is attenuated, and there is no evidence for the formation of multidimensional feature representations within the region (Wikenheiser et al., 2017). These data indicate that hippocampal input is not only necessary to mPFC coding of predicted outcomes, but also mPFC representation of the inferred multidimensional states that lead to those outcomes.

A number of recent human neuroimaging studies have further implicated mPFC in schema representations that index goal-relevant commonalities and differences acquired across multiple events. Like hippocampus, mPFC representations of overlapping memory elements become more similar to one another after learning (Schlichting et al., 2015), with consolidation increasing such learning-related representation of event commonalities within mPFC (Tompary & Davachi, 2017). Together with hippocampus, mPFC also represents temporal regularities extracted across multiple event sequences. For instance, in temporal community learning (Schapiro et al., 2013; Schapiro et al., 2016), individuals incidentally view a sequence of objects, in which the order of the objects is determined by an underlying structure with three temporal communities determined by their transition probabilities (**Figure 1.2G**). With learning, hippocampal representations become more similar for objects from the same temporal community and further distinguish members of different temporal communities, reflecting acquisition of the hierarchical temporal structure (Schapiro et al., 2016). Medial PFC responses also reflect learning of the structure, with increased engagement as participants view a sequence of objects from the same temporal community, suggesting its role in predicting what might be seen next in the sequence (Schapiro et al., 2013). Furthermore, mPFC—hippocampal connectivity is altered at the community boundaries, where a transition to a new temporal "state" is highly likely (Schapiro et al., 2013). Such findings provide initial speculative evidence for the role mPFC—hippocampal interactions in extracting the latent causes that represent multidimensional predictions about the environment.

More recently, multivariate analysis approaches have shown that hierarchical representations of task structures can be decoded from patterns of mPFC activation during well-learned tasks (Schuck et al., 2016) similar to work in rodents (Wikenheiser et al., 2017). These hierarchical representations reflected sixteen unobservable task states that predicted the possible sequences of actions participants could take to achieve a goal. Activation of a given task state during task performance within mPFC not only predicted when participants executed the correct choices, but was also predictive of their individual pattern of errors. Combining such representational neuroimaging approaches with computational modeling, Chan and colleagues (2016) further found that mPFC responses directly encoded information about inferred latent causes that represent how different combinations of task features are related to different outcome contexts (**Figure 1.2F**). When considered in light of other recent evidence for mPFC representation of common structure across real-world narratives (Baldassano et al., 2018), these findings provide compelling evidence for the role of mPFC in representing schematic information about latent causes.

In the same way hippocampal inputs to mPFC influence representation of the inferred latent causes of outcomes, top-down signals from mPFC may refine representation of hierarchical structure within hippocampus during schema formation. The mPFC—hippocampal circuit can thus be thought of as a dynamic loop that works together to organize knowledge in the most adaptive way possible (**Figure 1.1**). Specifically, mPFC may bias hippocampal encoding processes to emphasize the most goal-relevant features of the environment, while compressing irrelevant features that are not predictive of task outcomes. Lesion studies indicate that the mPFC plays a critical role in allocating attention to predictive task attributes during both decision making and subjective valuation (Vaidya & Fellows, 2015, 2016; Vaidya et al., 2018).

Recent model-based neuroimaging studies have provided further evidence that mPFC attentional processes influence extraction of latent causes and shape representation within hippocampus (Mack et al., 2016; Mack et al., 2020).

Specifically, in these studies, participants learned to classify insects into categories based on different combinations of predictive features. Some categorization problems required attention to a single feature, and others a combination of two or three features (**Figure 1.3A**). Successful learning thus required individuals to shift attention to the most diagnostic features for any given categorization problem. Across learning, mPFC representations extracted latent factors that captured the structure of the categories (Mack et al., 2020). Importantly, the dimensionality of the mPFC representations that emerged with learning tracked the problem complexity, with reduced dimensionality for simpler problems that required attention to fewer features and was directly related to participants' attentional strategies. Furthermore, mPFC—hippocampal interactions early in learning (**Figure 1.3B**) resulted in the formation of structured representations within hippocampus (**Figure 1.3A**), wherein category-relevant commonalities and differences were emphasized (Mack et al., 2016). These results suggest that the extraction of latent causes in mPFC may in turn shape how events are represented in hippocampus; associative features that are not reliably predictive of outcomes, and therefore not indexed by a latent cause, may ultimately be pruned from hippocampal representations (Kim et al., 2014; Kim et al., 2017).

5. Neural mechanisms supporting schema expression

Medial PFC—hippocampal interactions may be just as critical to expression of schemas once learned. When new situations overlap with existing knowledge, hippocampal pattern completion results in replay of related experiences within the circuit (Foster & Wilson, 2006; Wikenheiser & Redish, 2013). Replay events within CA_1 in particular modulate spiking activity within mPFC, reflecting the transmission of information associated with the current context to mPFC (Jadhav et al., 2016; Tang et al., 2017; Wang & Ikemoto, 2016). Furthermore, the memory content transferred from hippocampus to mPFC is highly structured, reflecting the detailed information about associative commonalities and differences (Tang et al., 2017).

Medial PFC may use this information to guide learning and decision making in the new context. For instance, mPFC activity is enhanced when new events can be processed in terms of an existing schema, facilitating subsequent memory for the new content (Tse et al., 2011; van Kesteren et al., 2013). When mPFC is damaged, however, schema-facilitated learning effects are eliminated (Spalding et al., 2015) highlighting the necessity of mPFC for the flexible expression of knowledge. In particular, mPFC may use reinstated associative information from hippocampus to generate predictions about which latent causes are most related to the current context; from those predictions, the mPFC may guide selection of the most adaptive set of actions to take to achieve one’s desired goal. Consistent with this hypothesis, patients with mPFC lesions have difficulty generating options to solve problems in real-world scenarios, such as having lunch at a restaurant and forgetting one’s wallet at home (Peters et al., 2017). Both in terms of the number of options generated and in the effectiveness of generated options, patients perform less well than healthy controls in deploying schematic knowledge to efficiently achieve their goals.

Reactivation of potential latent causes by mPFC may further refine retrieval of associative memory content within hippocampus. A large body of evidence indicates that PFC plays a specialized role in controlling hippocampal memory retrieval (Moscovitch, 1992; Simons

& Spiers, 2003). Medial PFC in particular may guide hippocampal retrieval in a similar manner to its influence on hippocampal encoding, by biasing retrieval toward the memories most relevant to the current context and suppressing activation of irrelevant content (Eichenbaum, 2017; Preston & Eichenbaum, 2013). Medial PFC—hippocampal interactions are enhanced when rodents stop at critical choice points in foraging tasks to consider the paths that might lead to a potential goal (Redish, 2016), with hippocampal responses reflecting forward replay of possible trajectories (Wikenheiser & Redish, 2015). Through the recent discovery of a monosynaptic pathway from mPFC to hippocampus in mice, researchers have directly shown that mPFC activity controls which hippocampal ensembles are active during memory retrieval and thus promotes context-specific retrieval of associative content within hippocampus (Rajasethupathy et al., 2015).

Additional evidence for the dynamic interactions between mPFC and hippocampus during context-dependent memory retrieval comes from the task depicted in **Figure 1.2A**. When rodents first entered one of the two task contexts, hippocampus relayed information about context-relevant associations to mPFC (Place et al., 2016). However, as the rodents approached the objects to make a choice, the flow of information reversed such that mPFC activity led the hippocampus. Critically, when mPFC is inactivated in this task, hippocampal population responses during object sampling no longer distinguish which options lead to reward in a given context, resulting in less accurate choices (Navawongse & Eichenbaum, 2013). Collectively, these results indicate that mPFC exerts top-down control of memory reinstatement in hippocampus in service of optimal decision making.

6. Neural mechanisms supporting schema updating

When schemas are flexibly expressed in new situations or learning contexts, there is often a need to update schemas to account for new information, both when it is consistent with predictions derived from schemas and when schema predictions are violated. In this section, we revisit Piaget's concepts of assimilation and accommodation, reviewing mechanistic evidence for the role of the hippocampus and mPFC in these two proposed processes. First, we begin with assimilation, which proposes that information congruent with schema predictions is rapidly incorporated into existing knowledge.

In a seminal set of studies, Tse and colleagues (2007; 2011) demonstrated the necessity of both hippocampus and mPFC in assimilation. In these experiments, rodents learned multiple flavor-location pairings within a spatial arena (**Figure 1.3C**). Over the course of several training sessions, rodents were placed in the arena and cued with one of the six flavors. Over time, the rodents increased their digging at the spatial locations associated with cued flavors, demonstrating learning of the initial spatial schema. The critical manipulation in these studies was the introduction of two new flavor-location pairings within the same environment (**Figure 1.3D**). While the rodents required several weeks to learn the initial flavor-location pairings, the two new pairings were learned within a single training session, providing behavioral evidence for their rapid assimilation into the existing spatial schema (Tse et al., 2007).

Critically, the researchers showed that the hippocampus was necessary for assimilating new pairs into the existing structure. Lesions to the hippocampus performed 24 hours after the introduction of the new pairs did not impact rodents' memory for either the original or newly learned pairs, further suggesting their rapid assimilation into cortical memory stores (**Figure 1.3E**). However, hippocampal lesions did prevent the rodents from incorporating additional new pairs into the spatial schema, indicating that hippocampus plays a key role in binding new events with existing memories. Moreover, follow up work with the same paradigm (Tse et al., 2011) showed that mPFC coordinates with hippocampus during schema assimilation. During the introduction of the new pairs, hippocampal-dependent learning was further supported by upregulation of immediate early genes in mPFC. Pharmacological inactivation of the mPFC disrupted the retrieval of both the originally learned spatial schema and new pairs (**Figure 1.3F**), and also disrupted acquisition of new pairs within the same environment. These results demonstrate a critical role for hippocampus and mPFC in schema expression and assimilation, as well as how schema knowledge has a facilitative effect on new learning.

Schema updating may also occur through accommodation when new events signal an existing schema is no longer a valid representation of environmental contingencies. In this case, the existing schema may be altered, or a new schema created to account for changes in the environment. As with initial learning, schema modification may be supported by mPFC—hippocampal interactions as indicated by the recent work in category learning discussed previously (Mack et al., 2016; Mack et al., 2020). In the category learning task depicted in **Figure 1.3A**, participants learn to categorize the insects according to one rule before the classification problem changes, requiring participants learn a new category structure for the same set of stimuli. In these studies, participants learn to categorize the insects according to three distinct rules in succession. Thus, the task demands require that participants update their category schemas to account for changes in the problem rules. During the periods when the category learning problem switches, increased connectivity between the mPFC and hippocampus is observed (**Figure 1.3B**). Moreover, hippocampal representations themselves are altered, rapidly changing how they represent the same stimuli after a problem switch (**Figure 1.3A**). Specifically, the hippocampus reorganizes the pattern of similarities and differences to focus organization on the features that are newly diagnostic of category membership (Mack et al., 2016). These findings provide direct evidence for how hippocampal representations themselves are altered through accommodation.

Schema updating is thought to directly rely on prediction error signals (Schlichting & Preston, 2015). When new events differ from reactivated memories, hippocampal engagement, and CA_1 activity in particular (Chen et al., 2011; Zeithamova et al., 2016), increases, signaling when predictive associative relationships have changed (Kumaran & Maguire, 2009; Olsen et al., 2012). Prediction error can promote the modification of hippocampal associative representations, either directly (Kim et al., 2014; Kim et al., 2017) or possibly through interactions with mPFC (Dunsmoor et al., 2019). Moreover, prediction error signals may promote memory updating in a dose dependent manner. When prediction errors are large, new schemas may be created. However, when prediction errors are smaller, schema updating may occur.

For instance, after learning of an association between a tone and a shock, extinction is more effective when extinction trials (presentation of a tone in the absence of a shock) are introduced gradually, interspersed with presentation of tone-shock pairings and gradually increasing in frequency until only extinction trials are presented (Gershman, Jones, et al., 2013). Spontaneous recovery of fear memories is less likely under such conditions, relative to the case when the schedule is reversed such that extinction trials are more frequent at the beginning of extinction learning and tone-shock pairings are more frequent at the end. In the gradual case, smaller prediction errors may promote updating of the original fear memory, thus permanently reducing the fear response. In contrast, large prediction errors elicited when extinction trials are concentrated early in the trial sequence may lead to formation of an extinction memory that is separate from the initial fear memory, resulting in the spontaneous recovery of fear at later timepoints. Thinking in terms of latent causes, the massed extinction training may be assigned to a different latent cause from the original tone-shock pairings, whereas the gradual extinction case may update the associative predictions indexed by an existing latent cause. Hippocampal prediction errors in particular may help drive updating of latent causes represented in mPFC (**Figure 1.1**), determining when existing schemas are updated versus when new schemas are created.

7. Summary of framework for the neural instantiation of schemas

Schema theory has deep historical roots in psychological research, which has long demonstrated the influence of structured knowledge on behavior. Although early schema theory raised fundamental questions regarding how schematic knowledge is formed and updated, the inherent difficulty of measuring these complex representational structures in situ has impeded our understanding of the precise mechanisms involved. Moreover, these measurement challenges have contributed to the longstanding controversy regarding the precise definition of the term "schema." The relatively recent introduction of new representational analysis techniques has since afforded direct measurement of neural schemas as they are formed and later expressed. In the present chapter, we thus provide a testable framework that explains the psychological principles associated with schemas in more mechanistic, neural terms. In particular, we propose a representational account that defines schemas in terms of how the brain represents overlapping experiences, thus beginning to address the "vexing" question of how a schema is learned. In light of the growing body of neuroscience work on the topic, we argue that the mPFC—hippocampal circuit serves as a dynamic loop that works together to extract and organize schematic knowledge (**Figure 1.1**). We further situate this operational definition with respect to the historical principles that guide schema theory, highlighting how this dynamic circuit captures the most goal-relevant features of the environment. The proposed neural framework provides a strong foundation from which to understand the psychological effects of schemas on behavior.

The present neural framework describes how initial schemas are formed, how they are expressed in new situations, and how they are updated over time to represent the complex structure of the environment, providing a comprehensive theory of schemas. As reviewed here, extensive neurocomputational evidence indicates that hippocampus and mPFC are both necessary to schema representation, with their interactions mediating schema formation by preferentially representing commonalities and differences across individual experiences (see **Figure 1.1**). Specifically, hippocampal pattern completion mechanisms trigger reactivation of events that overlap with new experiences. Hippocampus may then bind together events that share common features and outcomes as well as differentiate events that result in different outcomes. Importantly, hippocampus indexes associative links that are both directly experienced and extracted across multiple events.

The mPFC then uses associative input from hippocampus to extract the latent causes that represent the complex multidimensional structure of the environment. Thus, the neural instantiation of schemas may itself be hierarchical, with hippocampus serving as an index of associative links between distributed neocortical patterns, and mPFC indexing the combination of associative links (represented in hippocampus) that predict similar outcomes. Damage to either of these brain regions would thus exert a profound impact on schema formation and access. Medial PFC input to the hippocampus may further guide the refinement and updating of schemas, emphasizing representation of goal-relevant commonalities and differences in hippocampus, while deemphasizing idiosyncratic details that are not generally predictive. Similarly, interactions between these structures mediate retrieval of schemas once formed to guide selection of the most adaptive set of actions to take in new settings. Taken together, through specifying the mechanisms that support formation, expression, and updating of associative knowledge structures, the current chapter provides an overarching framework for how the fundamental properties of schemas may be understood through and implemented by neural mechanisms.

7.1 Contributions to schema theory: A new avenue for pursuing testable predictions

How does the proposed framework advance theoretical models of schemas? While the extant schema literature has demonstrated the profound effects that knowledge exerts on behavior, the lack of mechanistic clarity has limited the field's ability to move beyond description to systematically test predictions derived from theoretical accounts. For example, a prominent prediction of schema theory is that schemas support flexible reasoning behavior. Yet without a concrete method of measuring a schema, failure to demonstrate a hypothesized behavioral effect may be attributed to a failure to invoke the hypothesized schema, with no way of falsifying the prediction within a given experimental context. As Taylor and Crocker (1981) argued in their seminal review on the topic, until the theoretical link between experience, schemas, and behavior is formalized in a falsifiable form, the heuristic value of schema theory will be limited.

In the current chapter, we acknowledge the heuristic limitations of schema theory, and thus advocate for a formal, neural account that enables the field to test the underlying principles in falsifiable ways. First, if individuals construct schemas that capture goal-relevant features of the environment as Tolman predicted, then their representational structure should reflect this organization. Consistent with this prediction, recent model-based fMRI work has directly compared the neural representations elicited during tasks that exhibit different relational structures while controlling the individual features encoded. Finding that representations are compressed for simpler relational task structures relative to more complex relational tasks provides direct support for the theoretical prediction that schema representations capture goal-relevant features of the environment.

Second, if activation of existing schemas exerts a direct influence on learning and inference as Bartlett proposed, the degree of schema reinstatement should predict subsequent behavior. Indeed, as highlighted throughout this chapter, the introduction of neural decoding techniques has taken a critical step forward in establishing a direct, positive association between the degree of schema retrieval and a host of flexible behaviors, including reasoning and decision making. Finally, if Piaget's proposal that schemas are updated and altered to incorporate new experiences is true, experiences that are inconsistent with prior schemas should elicit modifications to underlying representations and/or the formation of new schemas. As discussed above, new research in humans and animals has provided initial insights into how schemas are updated by elucidating how prediction errors modify existing memories.

As these examples demonstrate, the mechanistic approach offered here, and afforded by modern neuroscience techniques, serves to complement and extend current schema theory, providing new avenues for testing and refining the core predictions. Importantly, we do not claim to have answered the longstanding question of how to define a schema. Nor do we argue that the simplified tasks reviewed here necessarily constitute the types of complex schemas examined in the cognitive literature. However, through tightly controlled investigations of what relational structures are evoked by certain experiences and how reinstatement of those structures influences behavior, we may begin to develop a systematic understanding of the nature and function of these powerful mental structures. In accordance with this view, we note that Bartlett shared this general perspective 90 years ago when he argued that if it were possible to forgo “a single descriptive word” (i.e., schema) that “it would probably be best to speak of active, developing patterns” (Bartlett, 1932). With the introduction of neuroscientific techniques capable of measuring these patterns in situ as schemas are dynamically formed, updated, and expressed, we believe that this approach offers a complimentary path forward, lending explanatory power to current schema theory.

8. Future Directions

The findings reviewed here add to a growing body of evidence that dynamic mPFC—hippocampal processing and representation support a host of cognitive functions, providing efficient access to relevant knowledge during memory, inference, and decision-making. Because the individual elements of experience are thought to be distributed throughout the neocortex (Teyler & DiScenna, 1986), it is possible that the hippocampus and mPFC act as hierarchical indices that route behaviorally-relevant information to sensory-specific cortical areas to support behaviors in new situations (Mack & Preston, 2016). That is, we speculate that a schema may not be stored in any one region of the brain, but rather consists of hierarchical representations in sensory-specific cortical regions, with hippocampus and mPFC serving as indices to these distributed cortical patterns. The research outlined here may thus suggest a broader theoretical view of the functional role of the hippocampus and mPFC in supporting complex forms of cognition. Future work is needed to test this prediction directly.

Another key question about schemas is the level of detail that may be represented. The bulk of the cognitive neuroscience work reviewed here suggests that associative structures come to represent overlapping commonalities while also preserving the elements of individual events (e.g., Schlichting et al., 2015). In contrast to this view, other perspectives proffer that schemas consist of only the abstracted elements of experience, arguing that the loss of idiosyncratic details is an essential component of a schema (Ghosh & Gilboa, 2014). Although the hippocampal—mPFC circuit engaged during the tasks reviewed here share many of the properties of schemas, it will be left to future research to determine whether this framework functions similarly in more traditional schema tasks which have defined it solely as an abstracted mental structure. Furthermore, though many of the associative learning tasks described here facilitate disentangling the basic building blocks of associative knowledge formation, they do not allow for examination of the complex features of real-world schemas. An important avenue for future research thus entails examination of whether the neural circuit examined here accounts for schemas that scale in complexity, both in terms of content and structure.

A final question ripe for future research concerns what role time plays in schema formation and retrieval. The original work by Bartlett found that schema-induced interference increased over time as individuals lost the details of individually encoded events, suggesting that these representations are fundamentally altered with consolidation. Yet to our knowledge, only one study to date has measured schemas over extended time periods in humans (Sommer, 2017). Over the course of one year, repeated retrieval of associative structures shifted from the hippocampus to the PFC. It has been proposed that, over time, memories lose their association to the original context thereby becoming semanticized (O'Reilly et al., 2014). In light of the neural framework proposed here, memory consolidation may cause accelerated forgetting of idiosyncratic information while also strengthening goal-relevant commonalities and differences. Future research aimed at clarifying how these associative structures transform over time will provide additional insight into the mechanisms involved in schema formation and transformation. In addition to mapping the trajectory of schematic change within an individual, mechanistic insight into schemas will further be gained from systematic investigations of how the neurobiological underpinnings of schemas change over the course of development. Because the hippocampus and mPFC continue to develop through adolescence, improvements in reasoning may thus be rooted in the development of neural structures that support schema formation and expression.

In conclusion, early psychological theories of schemas have recently been augmented by a growing neuroscientific literature that clarifies the mechanisms involved in schema formation, updating, and expression. By quantifying schemas as “active, developing patterns” of neural activity, neuroscientific work in animals and humans has advanced schema theory by documenting how schematic knowledge is organized in neural terms and used to guide behavior in both familiar and new contexts. The deployment of new representational analysis approaches to neural data has thus opened a new door for research on schemas, providing several open avenues for future research, which may help to resolve many of the longstanding challenges for schema theory.

Box 1.1. Bayesian inference guides decisions under uncertainty

In some cases, the relevant schema for a situation may be obvious based on contextual cues. For example, seeing a terminal with shops and gates, with planes docked at them, provides a clear signal that you are in an airport. However, because schemas are based on latent causes, in practice the relevant schema may not be immediately clear. In these cases, one must use available evidence to infer which schemas are mostly like to apply to the current situation. For example, if one is meeting a friend at an unfamiliar restaurant, one might not know initially whether it is a casual restaurant or a formal restaurant with a dress code (**Box Figure 1.1A-B**). Based on one's previous experience learning about casual and formal dining contexts, one might know that, in casual contexts, one is more likely to see someone wearing jeans than someone wearing a suit. Conversely, formal contexts are more associated with suits than with jeans. When entering this new restaurant, seeing what individual patrons are wearing provides evidence about what kind of restaurant this is (**Box Figure 1.1C**).

Bayes' theorem defines the optimal way to update your beliefs to take into account both prior experience and new evidence (**Box Figure 1.1D**). In Bayes' theorem, beliefs are expressed as probabilities; for example, before entering an unfamiliar restaurant, one might believe there is a 40% chance that it is a formal restaurant. According to Bayes' theorem, prior beliefs (e.g., the a priori probability that a given restaurant will be formal) should be updated to reflect relevant observations (e.g., seeing a patron wearing a suit), forming what is known as a posterior probability that reflects one's updated belief based on the new evidence. Because this calculation makes it possible to reason about possible situations in the absence of conclusive evidence, it provides a powerful method to make inferences about latent causes. A recent study found that mPFC represents the posterior probability of different potential, well-defined latent causes (Chan et al., 2016; **Figure 1.2F**), suggesting that mPFC may be involved in computing something similar to Bayesian inference.

However, exact Bayesian inference requires evaluating every possible latent cause that may be currently relevant, which is often infeasible. For example, when walking into a new restaurant context, one might be surprised to see a pool table, because this observation was unlikely under all of the different restaurant schemas that you have. According to the latent cause model (Gershman et al., 2017), there are two ways to respond to a situation that is not well explained by previously known latent causes. Faced with this unexpected outcome, one could update knowledge of restaurants to accommodate this new information that a restaurant can also include games like pool. Alternatively, one could decide that this new establishment is a different sort of context that cannot be accounted for in the restaurant schema, and thus attribute observations about this context to a different latent cause. In this case, a new schema would be formed, and the existing restaurant schema would not be modified. The latent cause model proposes that prior beliefs help to determine whether two events are due to the same latent cause or different latent causes. The model proposes that, all else being equal, two events that are nearby in time will tend to be related to the same latent cause. This prior belief makes temporally and spatially contiguous events more likely to be assumed to be related to the same latent cause (Gershman et al., 2017; Soto et al., 2014). The same mechanism may also inhibit integration of events that are temporally or spatially distant from one another (Cai et al., 2016; Rashid et al., 2016; Zeithamova & Preston, 2017). More generally, multiple features of events, such as temporal, spatial, or conceptual distance (Morton et al., 2017) may influence one’s prior beliefs about whether two events have the same latent cause, influencing whether those events become integrated or separated within schematic knowledge.

References

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin, 93*(2), 203-231.

Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *J Neurosci, 38*(45), 9689-9699.

Barbas, H., & Blatt, G. J. (1995). Topographically specific hippocampal projections target functionally distinct prefrontal areas in the rhesus monkey. *Hippocampus, 5*(6), 511-533.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, England: Cambridge University Press.

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learn Mem, 11*(5), 485-494.

Bunsey, M., & Eichenbaum, H. (1996). Conservation of hippocampal memory function in rats and humans. *Nature, 379*(6562), 255-257.

Cai, D. J., Aharoni, D., Shuman, T., Shobe, J., Biane, J., Song, W., Wei, B., Veshkini, M., La-Vu, M., Lou, J., Flores, S. E., Kim, I., Sano, Y., Zhou, M., Baumgaertel, K., Lavi, A., Kamata, M., Tuszynski, M., Mayford, M., Golshani, P., & Silva, A. J. (2016). A shared neural ensemble links distinct contextual memories encoded close in time. *Nature, 534*(7605), 115-118.

Cavada, C., Company, T., Tejedor, J., Cruz-Rizzolo, R. J., & Reinoso-Suarez, F. (2000). The anatomical connections of the macaque monkey orbitofrontal cortex. A review. *Cereb Cortex, 10*(3), 220-242.

Chan, S. C., Niv, Y., & Norman, K. A. (2016). A probability distribution over latent causes, in the orbitofrontal cortex. *J Neurosci, 36*(30), 7817-7828.

Chen, J., Olsen, R. K., Preston, A. R., Glover, G. H., & Wagner, A. D. (2011). Associative retrieval processes in the human medial temporal lobe: hippocampal retrieval success and CA1 mismatch detection. *Learn Mem, 18*(8), 523-528.

Collin, S. H., Milivojevic, B., & Doeller, C. F. (2015). Memory hierarchies map onto the hippocampal long axis in humans. *Nat Neurosci, 18*(11), 1562-1564.

Deuker, L., Bellmund, J. L., Navarro Schroder, T., & Doeller, C. F. (2016). An event map of memory space in the hippocampus. *Elife, 5*.

Dunsmoor, J. E., Kroes, M. C. W., Li, J., Daw, N. D., Simpson, H. B., & Phelps, E. A. (2019). Role of human ventromedial prefrontal cortex in learning and recall of enhanced extinction. *J Neurosci, 39*(17), 3264-3276.

Dusek, J. A., & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Sciences U S A, 94,* 7109-7114.

Eichenbaum, H. (2017). Memory: Organization and control. *Annu Rev Psychol, 68,* 19-45.

Farovik, A., Place, R. J., McKenzie, S., Porter, B., Munro, C. E., & Eichenbaum, H. (2015). Orbitofrontal cortex encodes memories within value-based schemas and represents contexts that guide memory retrieval. *J Neurosci, 35*(21), 8333-8344.

Foster, D. J., & Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature, 440*(7084), 680-683.

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

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychol Rev, 117*(1), 197-209.

Gershman, S. J., Jones, C. E., Norman, K. A., Monfils, M. H., & Niv, Y. (2013). Gradual extinction prevents the return of fear: implications for the discovery of state. *Front Behav Neurosci, 7,* 164.

Gershman, S. J., Monfils, M. H., Norman, K. A., & Niv, Y. (2017). The computational nature of memory modification. *Elife, 6.*

Gershman, S. J., Schapiro, A. C., Hupbach, A., & Norman, K. A. (2013). Neural context reinstatement predicts memory misattribution. *J Neurosci, 33*(20), 8590-8595.

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia, 53,* 104-114.

Hindy, N. C., Ng, F. Y., & Turk-Browne, N. B. (2016). Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nat Neurosci, 19*(5), 665-667.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychol Rev, 93*(4), 411-428.

Jadhav, S. P., Rothschild, G., Roumis, D. K., & Frank, L. M. (2016). Coordinated excitation and inhibition of prefrontal ensembles during awake hippocampal sharp-wave ripple events. *Neuron, 90*(1), 113-127.

Ji, D., & Wilson, M. A. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat Neurosci, 10*(1), 100-107.

Johnson, A., & Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J Neurosci, 27*(45), 12176-12189.

Karlsson, M. P., & Frank, L. M. (2009). Awake replay of remote experiences in the hippocampus. *Nat Neurosci, 12*(7), 913-918.

Kim, G., Lewis-Peacock, J. A., Norman, K. A., & Turk-Browne, N. B. (2014). Pruning of memories by context-based prediction error. *Proc Natl Acad Sci U S A, 111*(24), 8997-9002.

Kim, G., Norman, K. A., & Turk-Browne, N. B. (2017). Neural differentiation of incorrectly predicted memories. *J Neurosci, 37*(8), 2022-2031.

Koscik, T. R., & Tranel, D. (2012). The human ventromedial prefrontal cortex is critical for transitive inference. *J Cogn Neurosci, 24*(5), 1191-1204.

Kuhl, B. A., Bainbridge, W. A., & Chun, M. M. (2012). Neural reactivation reveals mechanisms for updating memory. *J Neurosci, 32*(10), 3453-3461.

Kumaran, D., & Maguire, E. A. (2009). Novelty signals: a window into hippocampal information processing. *Trends Cogn Sci, 13*(2), 47-54.

Kumaran, D., Warren, D. E., & Tranel, D. (2015). Damage to the ventromedial prefrontal cortex impairs learning from observed outcomes. *Cereb Cortex, 25*(11), 4504-4518.

Lacagnina, A. F., Brockway, E. T., Crovetti, C. R., Shue, F., McCarty, M. J., Sattler, K. P., Lim, S. C., Santos, S. L., Denny, C. A., & Drew, M. R. (2019). Distinct hippocampal engrams control extinction and relapse of fear memory. *Nat Neurosci, 22*(5), 753-761.

Larkin, M. C., Lykken, C., Tye, L. D., Wickelgren, J. G., & Frank, L. M. (2014). Hippocampal output area CA1 broadcasts a generalized novelty signal during an object-place recognition task. *Hippocampus, 24*(7), 773-783.

Lisman, J. E., & Grace, A. A. (2005). The hippocampal-VTA loop: controlling the entry of information into long-term memory. *Neuron, 46*(5), 703-713.

Liu, Z. X., Grady, C., & Moscovitch, M. (2017). Effects of prior-knowledge on brain activation and connectivity during associative memory encoding. *Cereb Cortex, 27*(3), 1991-2009.

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proc Natl Acad Sci U S A, 113*(46), 13203-13208.

Mack, M. L., Love, B. C., & Preston, A. R. (2018). Building concepts one episode at a time: The hippocampus and concept formation. *Neurosci Lett, 680*, 31-38.

Mack, M. L., & Preston, A. R. (2016). Decisions about the past are guided by reinstatement of specific memories in the hippocampus and perirhinal cortex. *Neuroimage, 127*, 144-157.

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications, 11*(1), 1-11.

Manns, J. R., & Eichenbaum, H. (2006). Evolution of declarative memory. *Hippocampus, 16*(9), 795-808.

Marr, D. (1971). Simple memory: A theory for archicortex. *Philos Trans R Soc Lond B Biol Sci, 262*(841), 23-81.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol Rev, 102,* 419-457.

McKenzie, S., Frank, A. J., Kinsky, N. R., Porter, B., Riviere, P. D., & Eichenbaum, H. (2014). Hippocampal representation of related and opposing memories develop within distinct, hierarchically organized neural schemas. *Neuron, 83*(1), 202-215.

Morton, N. W., Sherrill, K. R., & Preston, A. R. (2017). Memory integration constructs maps of space, time, and concepts. *Curr Opin Behav Sci, 17,* 161-168.

Moscovitch, M. (1992). Memory and working-with-memory: A component process model based on modules and central systems. *J Cogn Neurosci, 4*, 257-267.

Navawongse, R., & Eichenbaum, H. (2013). Distinct pathways for rule-based retrieval and spatial mapping of memory representations in hippocampal neurons. *J Neurosci, 33*(3), 1002-1013.

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014). Complementary learning systems. *Cogn Sci, 38*(6), 1229-1248.

Olsen, R. K., Moses, S. N., Riggs, L., & Ryan, J. D. (2012). The hippocampus supports multiple cognitive processes through relational binding and comparison. *Front Hum Neurosci, 6*, 146.

Pajkert, A., Finke, C., Shing, Y. L., Hoffmann, M., Sommer, W., Heekeren, H. R., & Ploner, C. J. (2017). Memory integration in humans with hippocampal lesions. *Hippocampus, 27*(12), 1230-1238.

Peters, S. L., Fellows, L. K., & Sheldon, S. (2017). The ventromedial frontal lobe contributes to forming effective solutions to real-world problems. *J Cogn Neurosci, 29*(6), 991-1001.

Piaget, J. (1954). *The construction of reality in the child*. New York: Basic.

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human Development, 15*(1), 1-12.

Place, R., Farovik, A., Brockmann, M., & Eichenbaum, H. (2016). Bidirectional prefrontal-hippocampal interactions support context-guided memory. *Nat Neurosci, 19*(8), 992-994.

Preston, A. R., & Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Curr Biol, 23*(17), R764-773.

Rajasethupathy, P., Sankaran, S., Marshel, J. H., Kim, C. K., Ferenczi, E., Lee, S. Y., Berndt, A., Ramakrishnan, C., Jaffe, A., Lo, M., Liston, C., & Deisseroth, K. (2015). Projections from neocortex mediate top-down control of memory retrieval. *Nature, 526*(7575), 653-659.

Rashid, A. J., Yan, C., Mercaldo, V., Hsiang, H. L., Park, S., Cole, C. J., De Cristofaro, A., Yu, J., Ramakrishnan, C., Lee, S. Y., Deisseroth, K., Frankland, P. W., & Josselyn, S. A. (2016). Competition between engrams influences fear memory formation and recall. *Science, 353*(6297), 383-387.

Redish, A. D. (2016). Vicarious trial and error. *Nat Rev Neurosci, 17*(3), 147-159.

Rescorla, R. A. (1988). Pavlovian conditioning. It's not what you think it is. *Am Psychol, 43*(3), 151-160.

Restle, F. (1961). *Psychology of judgment and choice*. New York: Wiley.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382-439.

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Curr Biol, 22*(17), 1622-1627.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nat Neurosci, 16*(4), 486-492.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (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*(1711).

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus, 26*(1), 3-8.

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nat Commun, 6*, 8151.

Schlichting, M. L., & Preston, A. R. (2015). Memory integration: neural mechanisms and implications for behavior. *Curr Opin Behavioral Sciences, 1*, 1-8.

Schlichting, M. L., & Preston, A. R. (2016). Hippocampal-medial prefrontal circuit supports memory updating during learning and post-encoding rest. *Neurobiol Learn Mem, 134*, 91-106.

Schlichting, M. L., Zeithamova, D., & Preston, A. R. (2014). CA1 subfield contributions to memory integration and inference. *Hippocampus, 24*(10), 1248-1260.

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron, 91*(6), 1402-1412.

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. *Neuron, 60*(2), 378-389.

Simons, J. S., & Spiers, H. J. (2003). Prefrontal and medial temporal lobe interactions in long-term memory. *Nat Rev Neurosci, 4*(8), 637-648.

Sommer, T. (2017). The emergence of knowledge and how it supports the memory for novel related information. *Cereb Cortex, 27*(3), 1906-1921.

Soto, F. A., Gershman, S. J., & Niv, Y. (2014). Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychol Rev, 121*(3), 526-558.

Spalding, K. N., Jones, S. H., Duff, M. C., Tranel, D., & Warren, D. E. (2015). Investigating the neural correlates of schemas: Ventromedial prefrontal cortex is necessary for normal schematic influence on memory. *J Neurosci, 35*(47), 15746-15751.

Spalding, K. N., Schlichting, M. L., Zeithamova, D., Preston, A. R., Tranel, D., Duff, M. C., & Warren, D. E. (2018). Ventromedial prefrontal cortex Is necessary for normal associative inference and memory integration. *J Neurosci, 38*(15), 3767-3775.

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nat Neurosci, 20*(11), 1643-1653.

Suh, J., Rivest, A. J., Nakashiba, T., Tominaga, T., & Tonegawa, S. (2011). Entorhinal cortex layer III input to the hippocampus is crucial for temporal association memory. *Science, 334*(6061), 1415-1420.

Tang, W., Shin, J. D., Frank, L. M., & Jadhav, S. P. (2017). Hippocampal-prefrontal reactivation during learning is stronger in awake compared with sleep states. *J Neurosci, 37*(49), 11789-11805.

Taylor, S. E., & Crocker, J. (1981). Schematic bases of social information processing. In E. T. Higgins, C. P. Herman, & M. P. Zanna (Eds.), *Social cognition: The Ontario symposium*. Hillsdale, NJ: Erlbaum.

Teyler, T. J., & DiScenna, P. (1986). The hippocampal memory indexing theory. *Behav Neurosci, 100*(2), 147-154.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychol Rev, 55*(4), 189-208.

Tompary, A., & Davachi, L. (2017). Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron, 96*(1), 228-241 e225.

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P., & Morris, R. G. (2007). Schemas and memory consolidation. *Science, 316*(5821), 76-82.

Tse, D., Takeuchi, T., Kakeyama, M., Kajii, Y., Okuno, H., Tohyama, C., Bito, H., & Morris, R. G. (2011). Schema-dependent gene activation and memory encoding in neocortex. *Science, 333*(6044), 891-895.

Vaidya, A. R., & Fellows, L. K. (2015). Ventromedial frontal cortex is critical for guiding attention to reward-predictive visual features in humans. *J Neurosci, 35*(37), 12813-12823.

Vaidya, A. R., & Fellows, L. K. (2016). Necessary contributions of human frontal lobe subregions to reward learning in a dynamic, multidimensional environment. *J Neurosci, 36*(38), 9843-9858.

Vaidya, A. R., Sefranek, M., & Fellows, L. K. (2018). Ventromedial frontal lobe damage alters how specific attributes are weighed in subjective valuation. *Cereb Cortex, 28*(11), 3857-3867.

van Kesteren, M. T., Beul, S. F., Takashima, A., Henson, R. N., Ruiter, D. J., & Fernandez, G. (2013). Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: from congruent to incongruent. *Neuropsychologia, 51*(12), 2352-2359.

van Kesteren, M. T., Rijpkema, M., Ruiter, D. J., Morris, R. G., & Fernandez, G. (2014). Building on prior knowledge: Schema-dependent encoding processes relate to academic performance. *J Cogn Neurosci, 26*(10), 2250-2261.

van Kesteren, M. T., Ruiter, D. J., Fernandez, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends Neurosci, 35*(4), 211-219.

van Strien, N. M., Cappaert, N. L., & Witter, M. P. (2009). The anatomy of memory: an interactive overview of the parahippocampal-hippocampal network. *Nat Rev Neurosci, 10*(4), 272-282.

Wang, D. V., & Ikemoto, S. (2016). Coordinated interaction between hippocampal sharp-wave ripples and anterior cingulate unit activity. *J Neurosci, 36*(41), 10663-10672.

Wang, S. H., & Morris, R. G. (2010). Hippocampal-neocortical interactions in memory formation, consolidation, and reconsolidation. *Annu Rev Psychol, 61*, 49-79, C41-44.

Wikenheiser, A. M., Marrero-Garcia, Y., & Schoenbaum, G. (2017). Suppression of ventral hippocampal output impairs integrated orbitofrontal encoding of task structure. *Neuron, 95*(5), 1197-1207 e1193.

Wikenheiser, A. M., & Redish, A. D. (2013). The balance of forward and backward hippocampal sequences shifts across behavioral states. *Hippocampus, 23*(1), 22-29.

Wikenheiser, A. M., & Redish, A. D. (2015). Hippocampal theta sequences reflect current goals. *Nat Neurosci, 18*(2), 289-294.

Wikenheiser, A. M., & Schoenbaum, G. (2016). Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. *Nat Rev Neurosci, 17*(8), 513-523.

Wimmer, G. E., & Shohamy, D. (2012). Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science, 338*(6104), 270-273.

Witter, M. P. (2011). Connectivity of the hippocampus. In V. E. A. Cutsuridis (Ed.), *Hippocampal microcircuits* (pp. 5-26). Berlin: Springer.

Yu, J. Y., Liu, D. F., Loback, A., Grossrubatscher, I., & Frank, L. M. (2018). Specific hippocampal representations are linked to generalized cortical representations in memory. *Nat Commun, 9*(1), 2209.

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron, 75*(1), 168-179.

Zeithamova, D., Manthuruthil, C., & Preston, A. R. (2016). Repetition suppression in the medial temporal lobe and midbrain is altered by event overlap. *Hippocampus, 26*(11), 1464-1477.

Zeithamova, D., & Preston, A. R. (2017). Temporal proximity promotes integration of overlapping events. *J Cogn Neurosci*, 1-13.

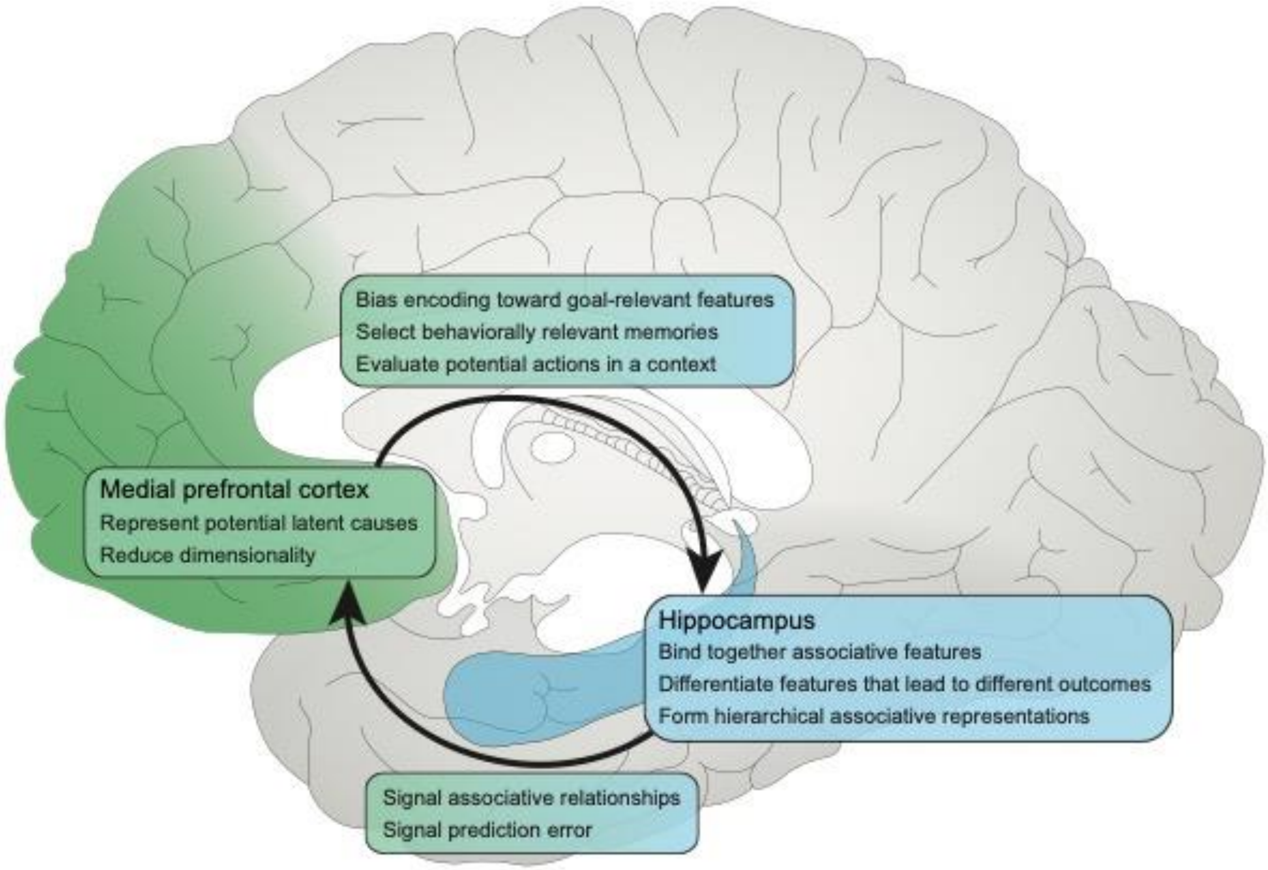

Bias encoding toward goal-relevant features
Select behaviorally relevant memories
Evaluate potential actions in a context
Medial prefrontal cortex
Represent potential latent causes
Reduce dimensionality
Hippocampus
Bind together associative features
Differentiate features that lead to different outcomes
Form hierarchical associative representations
Signal associative relationships
Signal prediction error

Figure 1.1. Schematic depiction of the hypothesized roles of hippocampus and medial prefrontal cortex (mPFC) and their connections in schema formation, expression, and updating. Hippocampal binding processes support formation of associations between event elements both within and across events. Such associative representations are thought to be hierarchically organized, emphasizing not only commonalities across events, but goal-relevant differences when outcomes vary by context. Hippocampal input to the mPFC provides information about associative relationships and may signal when new events differ from previously experienced events to promote schema updating. Medial PFC uses associative input from the hippocampus to extract information about the multidimensional associative features that predict the same or different outcomes, thus representing the likelihood of potential latent causes. Medial PFC is further thought to reduce the dimensionality of memory representations to emphasize encoding of features that are the most goal-relevant or predictive within hippocampus. Such dimensionality reduction may further support mPFC control of hippocampal retrieval, allowing for reinstatement of the most behaviorally relevant knowledge and evaluation of different potential actions during schema expression. Brain illustration courtesy of Margaret L. Schlichting.

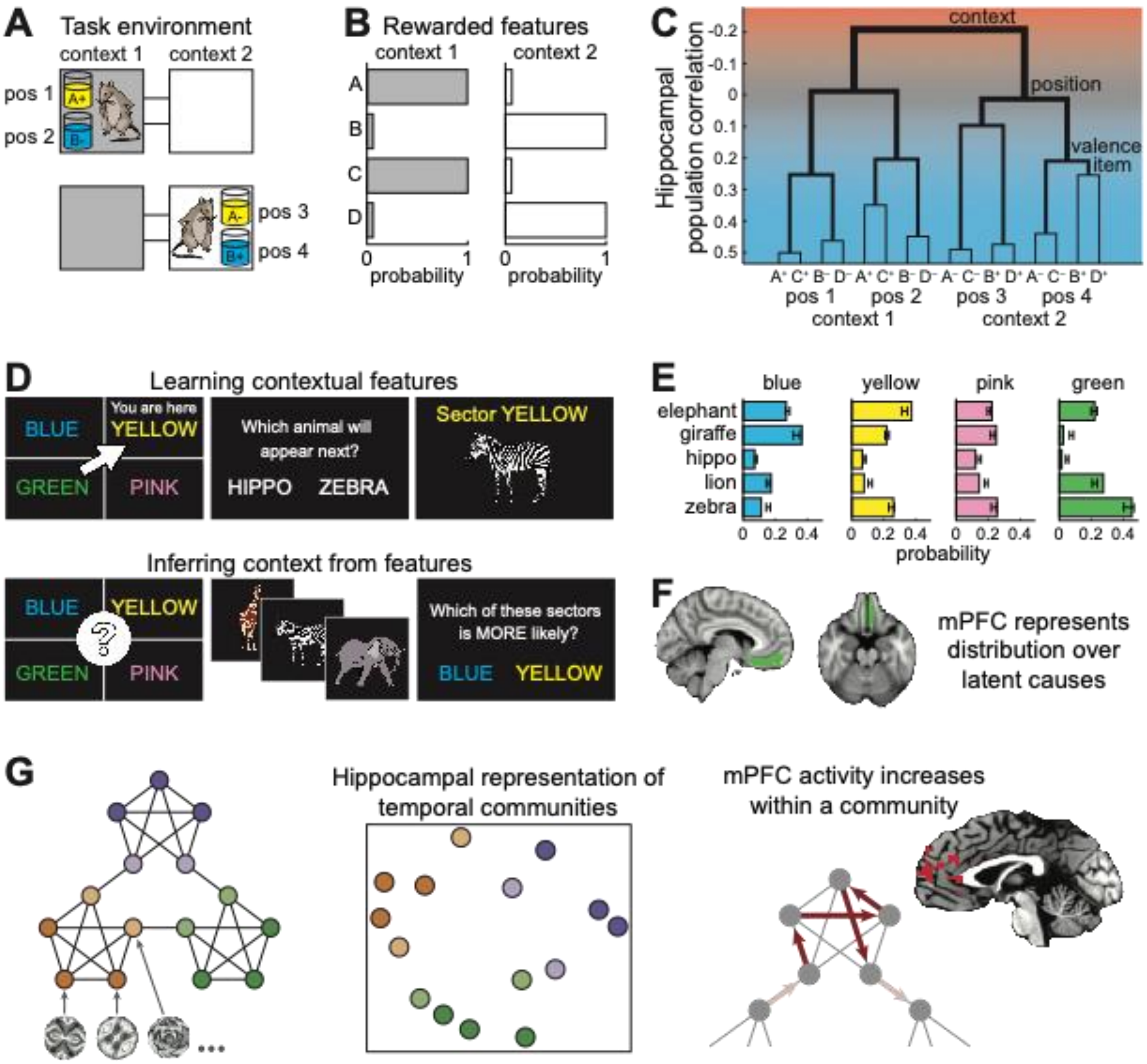


Figure 1.2. Hippocampus and mPFC form hierarchical representations that code the spatial, conceptual, and temporal contexts that lead to common or different outcomes. (**A**) Rodents learned to associate items (i.e., wells with scented digging material) with reward according to context-dependent rules. (**B**) In context 1, objects A and C were associated with reward, whereas in context 2, objects B and D were associated with reward. (**C**) After learning, hippocampal population responses demonstrated a hierarchical organization. When controlling for all other features, hippocampal responses showed anticorrelated activity (orange colors in the gradient) in the two contexts. Within the separate contexts, hippocampal responses reflected a hierarchical representation of similarities (blue colors in the gradient), with repetitions of the same item within the same context evoking the most similar responses, followed by valence, and then position (Panels A-C are adapted from McKenzie et al., 2014). (**D**) Learning about contextual associations allowed individuals to infer latent causes about where they were at within a global zoo context. Participants first learned about the probabilities of seeing individual types of animals (elephants, giraffes, hippos, lions, and zebras) in one of four sectors of the zoo (blue, yellow, pink, green). After learning, they were shown a sequence of animals and asked to infer

in which sector of the zoo they were located. (**E**) The actual probability distributions of seeing the five animals are depicted by the bars for each zoo context separately. The error bars reflect participants' estimates of the probability distributions, collected after the experiment. (**F)** The mPFC represented the probability distributions for each sector, conditional on the observed animals, suggesting its role in representing a distribution of possible latent causes (Panels D-F are adapted from Chan et al., 2016). (**G**) During temporal community learning, participants view a sequence of objects (fractal images in this example) while performing an incidental task. Unbeknownst to participants, the sequence is generated from a network structure that defines the transition probabilities between objects and has three distinct temporal communities comprised of five objects (here depicted as nodes in the purple, green, and orange communities). After learning, the patterns of hippocampal response evoked by members of the same temporal community are more similar than different community members. Medial PFC responses are also increased during "walks" through a single temporal community (depicted by red arrows in the structure), suggesting its sensitivity to the community structure (Panel G is adapted from Schapiro et al., 2013, 2016).

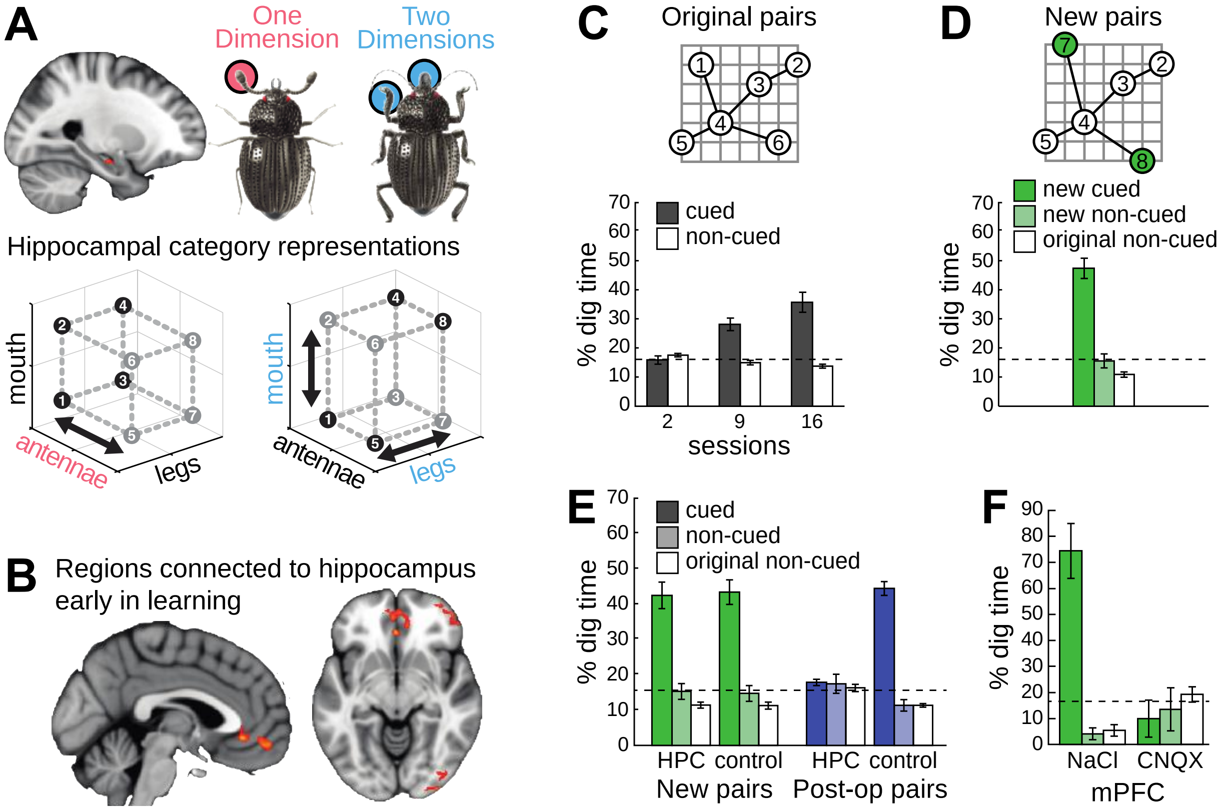
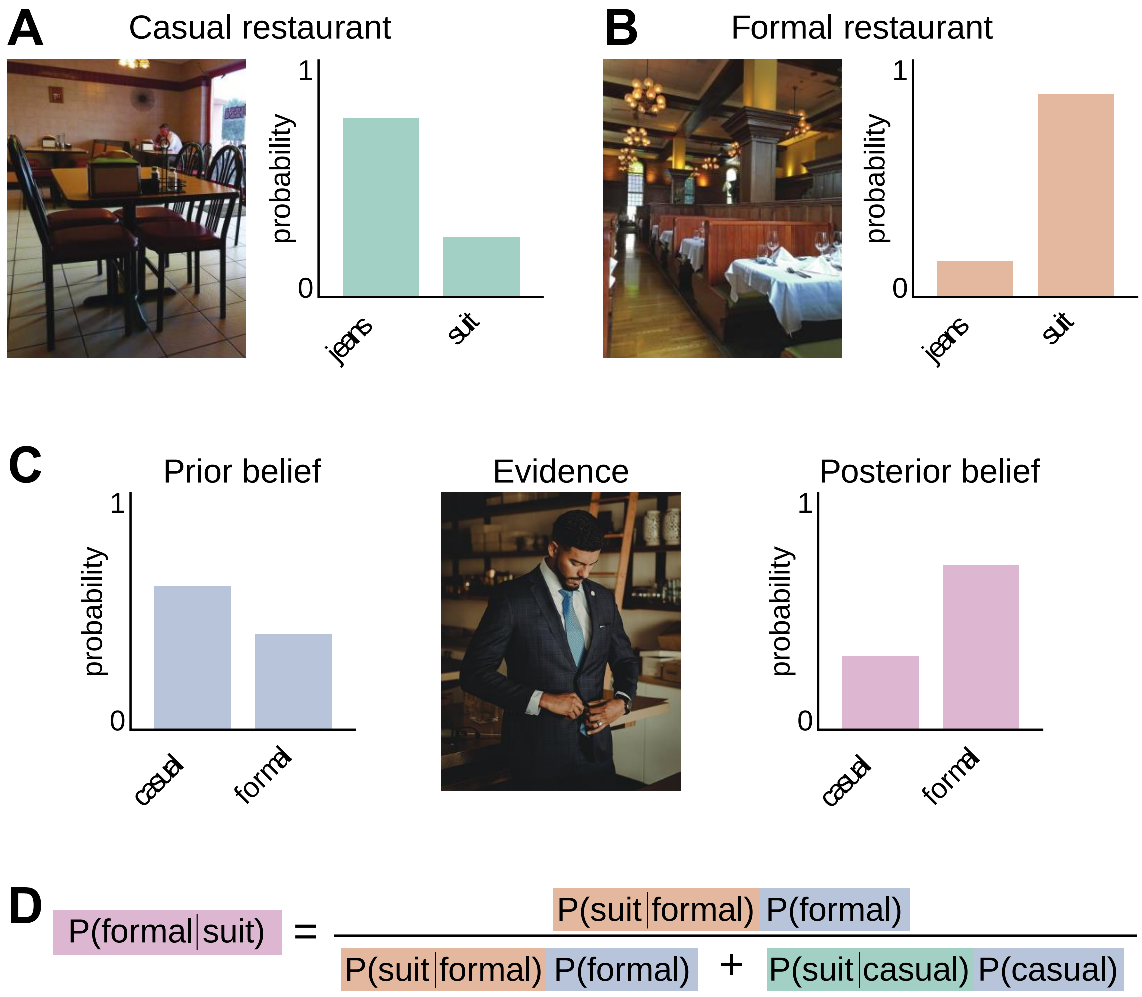


Figure 1.3. Hippocampus and mPFC support assimilation and accommodation. (**A**) Participants learned to categorize insects according to rules that differed in their complexity. Some category problems required attention to a single feature dimension, others two or three feature dimensions (the latter is not pictured). Hippocampal activation patterns reflected the optimal organization for category discrimination after learning and shifted when participants learned to categorize representations according to a new category rule, reflecting accommodation of the category schema. (**B**) Regions, including mPFC, showing increased connectivity with hippocampus early in learning as participants learned the category rules (Panels A-B are adapted from Mack et al., 2016). (**C**) Rodents learned a set of flavor-place associations (original pairs) within a spatial arena. Across initial training sessions, rodents increase their digging time in the spatial locations associated with cued flavors. (**D**) Rodents then updated their knowledge by learning new pairs in the same environment. While the original pairs were acquired over many sessions, rodents learned the locations of the new flavors in a single training session, suggesting that the new pairs were rapidly assimilated into an existing schema. (**E**) After 24 hours, hippocampal (HPC) lesions did not affect retrieval of the assimilated new pairs, but did prevent rodents from learning additional pairs (post-operative pairs) within the same environment. (**F**) Inactivation of mPFC (with CNQX) further prevented rodents from retrieving the assimilated new pairs relative to a saline (NaCl) control. (Panels C-F are adapted from Tse et al., 2007, 2011).



Box Figure 1.1. Bayes' theorem specifies how beliefs about the world, described in terms of probabilities, should be updated based on new evidence. (**A**) In the context of a casual restaurant, a given patron is more likely to be wearing jeans than a suit. (**B**) In contrast, in a formal restaurant, seeing someone wearing a suit would be more likely. (**C**) When first entering an unfamiliar restaurant, one will have a baseline expectation, known as a prior belief, of whether it is a casual or a formal restaurant. One might then encounter new evidence such as seeing a restaurant patron wearing a suit. Based on this evidence, one should update one's belief to reflect that the restaurant is most likely formal with a dress code; this revised belief is called a posterior. (**D**) Bayes' theorem calculates the optimal posterior belief after taking into account prior beliefs and the new evidence. Photographs courtesy of Rusty Clark, Alan Light, and 1DayReview, licensed under CC BY 2.0.