Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Nat Neurosci. Author manuscript; available in PMC 2020 May 21.

Published in final edited form as:

Nat Neurosci. 2019 October ; 22(10): 1544–1553. doi:10.1038/s41593-019-0470-8.

Learning task-state representations

Yael Niv

Psychology Department and Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA, 08544

Abstract

Arguably, the most difficult part of learning is deciding what to learn about. Should I associate the positive outcome of safely completing a street-crossing with the situation "the car approaching the crosswalk was red" or "the approaching car was slowing down"? In this Perspective, we summarize our recent research into the computational and neural underpinnings of "representation learning" — how humans (and other animals) construct task representations that allow efficient learning and decision making. We first discuss the problem of learning what to ignore when confronted with too much information, so that experience can properly generalize across situations. We then turn to the problem of augmenting perceptual information with inferred latent causes that embody unobservable task-relevant information such as contextual knowledge. Finally, we discuss recent findings regarding the neural substrates of task representations that suggest the orbitofrontal cortex represents "task states," deploying them for decision-making and learning elsewhere in the brain.

The ubiquitous problem of representation learning

Imagine standing on a street corner and preparing to cross the street on your way home (Figure 1A). Even in the calmest of neighborhoods, your sensory systems will confront a staggering amount of information that may or may not be relevant for the decision of whether to go or to wait. Computationally, avoiding getting run over is daunting. Nevertheless, you can probably complete the street-crossing task successfully even while talking to a friend or mentally planning your afternoon. What allows our brains to make decisions in complex, multidimensional environments with such ease and efficiency?

We argue that the brain solves seemingly complex tasks by learning efficient, low dimensional representations that simplify these tasks. A useful task representation will focus on aspects of the environment that are critical to correct performance of the task, that is, it will include all factors that are causally related to the outcome of our actions, for instance, the speed and distance of the closest oncoming car. At the same time, the task representation will gloss over all other information: the colors of the cars, the shops across the street, etc. Ignoring input dimensions (color, shape) that are irrelevant for task performance and concentrating on the few dimensions that are critical (speed, distance) allows us not only to make rapid decisions, but also to generalize learning as widely as possible. That is, correctly ignoring irrelevant aspects of the environment will allow learning from one experience to inform decision making in other scenarios that share relevant features with the current experience and differ only in the irrelevant ones.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Efficient representations are task-specific. When crossing the street, you should represent and act upon the speed and distance of cars, whereas when hailing a taxi, you should represent the color of the car, and whether or not the medallion light is on. How does the brain construct a representation for each of our numerous tasks? Summarizing a decade of our own research, here we discuss findings that suggest that two processes are critical to learning task representations: selective attention to only the relevant observable aspects of a task [1,2], and augmentation of these with hidden (unobservable) aspects from memory [3,4]. The learned task state, we will then argue, resides in the orbitofrontal cortex [5,6], which relays to other brain areas a pared down task representation tailored for each decision.

Reinforcement learning as a framework for decision making

In recent years, ideas from the computational field of reinforcement learning (RL; [7,8]) have revolutionized the study of learning and decision-making in the brain. In RL, tasks are defined as a set of *states* with action-dependent transitions between them, and with rewards (that can also be zero or negative) for each state. The agent starts at some state of the environment and traverses states based on (potentially probabilistic) transitions, collecting reward (or paying costs) throughout. Tasks do not have unique state representations, for instance, Figure 1B,C details two alternative representations of the same street crossing task. In Figure 1B, the representation includes four states (ovals), each defined by a location. Figure 1C represents the location "south sidewalk" as two distinct states, depending on whether a car is approaching or not, thus expanding the representation to five states. As can be seen from the action-dependent transitions in the figure (arrows, with actions in rectangles and probabilities of each transition), and as we detail below, some representations are more useful than others. In particular, by using the representation in Figure 1C to select actions, one can ensure reaching home safely, by choosing A2 ("wait") while in state S1b (car approaching) until such as time as a transition to S1a (no car approaching) occurs, at which point one can choose A1 ("go"). In contrast, given the state representation in Figure 1B, there is no action policy that will get you home while ensuring you don't get hit by a car (and therefore hospitalized).

RL algorithms provide a host of asymptotically converging methods for learning optimal behavioral policies that will maximize reward and minimize punishment within a given state representation [7–8]. Several reviews have detailed the correspondence between these algorithms and putative neural substrates, as well as challenges to this framework as an explanation of learning and decision making by humans and other animals (e.g., [9–11]). Briefly, in model-free RL, most algorithms are variants on the idea of using prediction errors to learn from experience a *value* for each state (or for each action in each state; so-called state-action Q values [12]) that summarizes the sum of future rewards that can be expected if one is in this state (and takes a particular action). Actions are then chosen that have a higher value and thus are expected to ultimately lead to more reward. In model-based RL, instead of learning values from experience, one uses experience to learn a model of the task – the transition probabilities between states, and the probability of reward in each state (e.g., the diagram in Figure 1B). The internal model is then used in a mental simulation to calculate the expected long-term return from different action options, and to choose the most rewarding one [13]. Evidence suggests that the brain uses both model-free and model-based

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

RL [14–15], with dopamine-dependent learning in the dorsal parts of the basal ganglia implementing the former [16–17], and prefrontal-hippocampal-medial basal ganglia circuitry implementing the latter [18–22], perhaps aided by dopamine released from the noradrenergic locus coeruleus [23–24]. However, the strict division between the two algorithms in the brain, and their neural substrates, has recently been challenged [25].

Reinforcement learning relies on representation of tasks as sequences of states

Designing the correct state space for each task is critical in RL [26–28]. First, different state representations will lead to different optimal action policies, some of which may never lead to the desired goal. For example, if you represent the street crossing task as in Figure 1B, you will stay on the curb forever, whereas the representation in Figure 1C will get you home safely. Second, RL methods suffer from *the curse of dimensionality* whereby the time (or iterations of learning) needed to solve the task scales exponentially with the number of states of the task [29]. If the state representation in Figure 1 included all pixels in the visual field, there would be numerous alternatives for S1, and it would take very many street crossings to estimate the values of different actions for each of the states accurately. Thus, to make learning in RL simulations feasible, state representations are typically crafted by hand, by an expert, so as to accurately describe the task with as few states as possible [27,28] (for example, [30]). The overwhelming majority of RL implementations, whether used to play backgammon [31] or to model animal learning (e.g., [32]), assume a state space, and are not concerned with learning it (but see [28,33]). A question then remains regarding how do living agents know what to represent [28] in order to use neural RL to solve these (and other) tasks?

Work in the booming field of deep learning has been specifically focused on learning useful, reduced representations – often not in the service of RL, but rather for classification tasks (e.g., labeling of images or face recognition). The marriage of deep learning and RL (that is, learning representations that are useful for goal-directed policy learning) has led to exciting breakthroughs in artificial intelligence, such as mastering the game Go [34], and playing Atari games better than humans [35]. However, these implementations require millions of iterations to learn representations, and therefore do not provide insight regarding how humans and animals learn task representations [36]. Moreover, even deep learning networks cannot flexibly solve new tasks that they have not been specifically programmed to learn [37], to the extent that these require a different representation of the same inputs (e.g., the task of hailing a taxi in Figure 1D).

A useful state representation for RL can be derived from the true *causal structure of the task* —it should include all the environmental features that determine (causally) whether actions will lead to (long term) task-relevant outcomes (rewards and punishments), while any dimension or feature of the environment that is not causal to these outcomes can be ignored and generalized over [27]. Learning such a minimal representation from trial and error is not trivial due to the large number of features that are potentially relevant to each task [28,38]. These include not only all the currently observed features, but also past events and actions

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

that may be causally relevant to future outcomes, for example, the fact that you hurt your foot yesterday and thus can only walk slowly across the street in Figure 1A. Moreover, while the environment typically provides feedback (rewards or punishments) for actions, there is no direct feedback for the representations underlying these actions, making it harder to determine if our current state representation is suitable for the task at hand, or can be further improved. Nevertheless, the brain seems to learn appropriate representations for new tasks almost effortlessly and in few trials [39]. How is representation learning achieved in the brain of humans and animals? Below, we summarize findings from several studies that were aimed at elucidating basic principles of neural representation learning, making note of open questions where progress can be made in the near term.

Dimensionality reduction of task representations using attention

As a highly-simplified laboratory analogue of a task such as street-crossing, where only few dimensions of the environment determine the outcome for actions, we developed the *Dimensions Task* ([1–2]; Figure 2A). On each trial, the participant was asked to choose one of three visual stimuli defined along three dimensions (e.g., shape, color and texture), in order to obtain reward. Importantly, participants were instructed that only one dimension (e.g., shape) is relevant to determining rewards, and within this dimension one 'target' feature (e.g., circle) will lead to a reward 75% of the time while other features will lead to reward with only 25% chance. Every few (e.g., 30) trials, the relevant dimension and the target feature changed, and the participant was notified ('a new game is now starting'). This task, similar to the extradimensional/intradimensional set-shifting task in animals [40], is a multidimensional extension of the multiple-option choice task (n-armed bandit task) used in many neuroeconomics studies [e.g., 41–42], and a probabilistic version of the Wisconsin card sorting task [43], albeit with rapid, signaled, changes of relevant dimension. Our goal was to use this task, together with a host of computational models, to investigate in detail how humans learn by trial and error what is the relevant dimension for representing task stimuli.

Findings from the Dimensions Task have shown that participants employ attention mechanisms to create a lower-dimensional representation that constrains learning and decision making to dimensions that are deemed relevant for obtaining reward [1,2]. In particular, formal model comparison showed that participants do not learn as Bayesian ideal observers on the one extreme, nor do they apply RL to each stimulus as a whole on the other extreme (e.g., learning a value for a green triangle with stripes and a separate value for a green triangle with polka dots) [1]. Instead, participants learn the task using attention-weighted RL [2,28]: stimulus dimensions are selectively weighted by attention both when determining the value of each option, and when learning from prediction errors (Figure 2B). These weights can be seen as constructing a lower-dimensional task representation. For instance, if the weight for one dimension (e.g., color) is 1 and the other two weights are 0, the task representation will include only the color of the stimuli, ignoring shape and texture. Other weights will prioritize dimensions less exclusively, for instance strongly representing shape, and ignoring color for the most part, but not completely. In this way, attention dynamically generates a representation that focuses on the dimensions that are deemed most relevant to task performance at any given timepoint [44].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

This work has highlighted the involvement of the fronto-parietal attention network (including the intraparietal sulcus, precuneus and dorsolateral prefrontal cortex; [45–46]), classically associated with attentional control, in changing attention from one dimension of the task to another during learning [2], and therefore for switching between alternative task representations. BOLD signals in both value-sensitive areas in the dlPFC and prediction-error-sensitive areas in the ventral striatum showed significantly higher correlation with values and prediction errors (respectively) derived from the attention-weighted model as compared to a model that attends equally to all three dimensions ([2]; Figure 2C), further supporting the idea that task representations are dynamically modified using attention [44,47,48]. However, it is not yet clear how humans determine what to attend to, and how attention changes from trial to trial based on actions and outcome feedback (Box 1).

The role of inference and context in shaping task representations

Dimensionality reduction through selective attention addresses one aspect of the representation learning problem: the existence of extraneous information in sensory input. An opposite problem is that of missing data: not all task-relevant information is readily available. For instance, in some countries you should look to the other direction in search of oncoming traffic; still in others, to cross the street you should not wait for a gap in traffic but rather walk at a steady, slow pace, and cars will slow down and create a gap around you, though a crosswalk may look deceptively similar in all cases. More generally, context – location, time, phase of the task, as well as internal goals and past knowledge – is often critical to learning and decision making, but is not necessarily perceptually observable.

Put differently, a ubiquitous problem in learning a task representation is that externally similar situations can require different action policies in different contexts, and thus should be represented as separate states depending on context. Conversely, some seemingly different situations may be equivalent in terms of the causal structure of the task, and thus should be aliased into a single state to allow generalization [27–28,49–50]. In real-world tasks, incoming information is rarely delineated into trials with clear, punctate, stimuli that flash on and off to signal the current state, thus a fundamental problem in learning is determining which experiences should be considered together (i.e., represented as the same state, allowing learning from one experience to affect behavior in the other) or learned about separately [4,51]. For instance, in the example in Figure 1, the question is whether instances of standing on the south sidewalk should be delineated into two states based on the presence or absence of oncoming traffic, as in Figure 1C (when attempting to go home), or rather they can all be considered as a single state (as is the case if Figure 1D, when going to work, although in that case instances in which a taxi is approaching need to be represented as a separate state). Regardless of whether the context is observable or not (in the example in Figure 1, the task goal can be considered a context), it is clear that situations must be grouped into states differently for different contexts, and this context- or task-specific grouping must be learned.

Since no two experiences are exactly alike, one idea is that incoming information is clustered according to *similarity* (i.e., similar experiences are assigned to the same cluster), with each cluster effectively forming a state on which RL operates (Figure 3A). This

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

intuitive but powerful idea has been formalized within a statistically optimal theory (based on Bayesian inference) of how an animal or human might generate task states in their internal representation of the world, based on experience. The idea is that we start from a single state (cluster), and expand our representation as needed when new events differ substantially from what we have experienced so far [3,52]. In this way, the high-level contextual structure of a task can be learned.

Such clustering-based representation learning, together with traditional associative learning that treats each cluster as a distinct state (e.g., learning a single association or value based on all stimuli assigned to that cluster), can explain why it is difficult to change old associations. Consider the notorious ineffectiveness of extinction procedures in fear conditioning: after a rat is exposed to pairing of a tone and a shock such that the tone comes to elicit freezing behavior due to prediction of an upcoming shock, attempts to weaken the association between the tone and the shock by repeatedly presenting the tone without shocks reduce fear only temporarily. Studies demonstrate that the fear response returns over time ('spontaneous recovery') [53], with changes in context ('renewal') [54], or after a reminder shock ('reinstatement') [55]. Our framework explains the ineffectiveness of extinction as resulting from the animal's inference that extinction trials and acquisition trials are likely generated by different 'latent causes,' and thus should be clustered into separate states (Figure 3A) [3]. This idea provides a normative explanation for the long-held view that in extinction the animal learns a new 'tone→no shock' association, rather than modifying the old 'tone →shock' association [54].

Indeed, increasing the similarity between the acquisition phase and the extinction phase through gradual extinction can increase the effectiveness of extinction. Gershman and colleagues [56] used the principle of similarity-based representation learning to modify fear memories in rats by making the change from shock to no shock gradual (in contrast to the commonly used abrupt extinction paradigm; Figure 3B). Their goal was to prevent the rats from generating a new task state in the extinction phase, and thus cause the 'tone →no shock' experiences to influence the previously acquired 'tone→shock' association. Gradual extinction resulted in more persistent loss of fear, as measured by lack of reinstatement or spontaneous recovery of fear at test (Figure 3C). In human fear conditioning, a meta-analysis showed that individuals who extinguished a fear association faster showed more spontaneous recovery of fear [57]. This may be because those who showed faster extinction had more readily inferred a new latent cause in extinction. As a result, they attributed fewer extinction trials to the acquisition cause, leaving the original fear memory intact, as was evident in the later recovery of fear.

Similarity-based clustering as a general principle of representation learning

Clustering of experience according to similarity for the purposes of delineating task states seems to be a general phenomenon in learning, going beyond situations that involve strong aversive reinforcers. For instance, in a perceptual decision-making task [58], humans were asked to quickly estimate the number of circles on a screen (Figure 4A). The true number of circles was then revealed, such that participants could estimate the number of circles on subsequent trials by learning the mean number of circles in the task (the number of circles

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

was drawn from a Gaussian distribution, making the mean of previous trials the best estimate for the next trial). Unbeknownst to participants there were two types of trials signaled by different colors of the circles (e.g., blue circles in one trial type and green circles in the other; Figure 4B) and drawn from different Gaussian distributions (Figure 4A). The resulting number estimates (Figure 4C) suggested that when the distributions associated with the two trial types were similar, participants grouped the different colored trials together and averaged their statistics during learning, as if they all belonged to one color-blind state. In contrast, when the two means were dissimilar, participants seemed to separate trials into single-color states (as in Figure 4B), separately learning about circles of each color [58].

Beyond their effect on delineating learning, clusters of experiences that are blended together in one state can be viewed as comprising a single multi-event 'memory trace'. The boundaries of memory traces may therefore also reflect the process of representation learning: if the current experience is similar enough to previous experiences, it should be added to a previous memory trace (thereby modifying that memory, i.e., causing learning [59]), whereas a novel experience should spur the creation of a new memory trace. To test this, participants were presented with line segments that changed gradually throughout a block of the experiment. Critically, in half of the blocks, among the small trial to trial changes of the line segment, one larger change ('jump') was embedded. The authors hypothesized that this jump would cause a splitting of memory traces, thereby protecting the memories of the line segments in the beginning of the block from interference from the segments later in the block. Indeed, when tested for memory of one of the initial line segments in the block, participants were more accurate in blocks that involved a jump as compared to blocks that only involved gradual changes. Computational modeling suggested that in the jump condition, two memory traces were created whereas blocks with no jump resulted in a single memory trace [60]. The idea that latent-cause inference determines not only RL, but also the organization and modification of memories, can explain many extant phenomena in the literature on 'reconsolidation' of memory [61].

The neural substrates of representation learning

Lesion and developmental data in rodents suggest that the hippocampus plays an important role in determining when to create a new state (cluster) or update an old one [3,61–65]. The hippocampus is important for detecting novelty [66–67] and therefore may participate in computing and signaling the (dis-)similarity of different experiences. Animals with hippocampal lesions do not show context sensitivity of extinction [62] or other related conditioning phenomena such as latent inhibition [63], instead behaving as if extinction and acquisition trials are all clustered in one state. Animals with an under-developed hippocampus (e.g., young humans and rodents) also generalize learning widely [64–65,68–69], consistent with an over-simplified representation with a small number of states [3].

Other work implicates a different brain area – the orbitofrontal cortex (OFC) – in the representation of learned task states [70–71]. Many decision-making related functions have been attributed to the OFC, including response inhibition, somatic markers, and most dominantly: outcome and/or economic value expectancies [5,71]. However, lesions [72–75], inactivation [76], electrophysiological recording data [77–81], and fMRI findings [6,82–84]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

suggest that the OFC is critical for representing the current state of the task when this state is not immediately evident from sensory information and conveying this representation to the striatum for RL [5]. For instance, in an fMRI study in which human participants were asked to infer what quadrant of a safari they were in based on evidence of animals recently seen, the similarity structure of optimal location inferences over time correlated best with the similarity structure of activations in the OFC [83]. In electrophysiological recordings from OFC in an odor-guided go/no go task in rodents where different odors appeared in sequences that made the task equivalent to traversing a (virtual) T-maze, state similarity showed that OFC represented the detailed state structure of the task, above and beyond other quantities such as expected rewards that could also be extracted from the neural dynamics [81].

Both electrophysiological [77] and lesion data [5] suggest that even without a functioning OFC, the striatum has access to a state representation that is bound to external stimuli ('observable states'; although unobservable information about timing is also likely computed in the striatum [85]). In contrast, inferred states that incorporate internal information from, e.g., working memory, intended actions, or future goals, and are based on latent-cause inference as discussed above ('partially observable states') seem to require the OFC [5,86–87]. This can explain why decision making becomes more OFC-dependent as tasks rely more on inference processes necessary to determine the underlying hidden state of a task.

To test this hypothesis, Schuck and colleagues designed a task in which correct performance required partially observable states [6]. Human participants had to judge the age (old/young) of either a face or a house presented overlaid on each other, with the category to be judged determined by the previous trial (Figure 5A). The initial category to be judged was instructed. Thereafter, participants were asked to continue judging that same category as long as the age of the stimuli in the judged category remained the same. Once the age of the currently-judged category differed from the age in the previous trial, they were instructed to switch to judging the alternative category on the next trial. Participants performed the task well (<5% errors) suggesting that they correctly represented, at each timepoint, the state of the task (Figure 5B). A multivariate classifier for the 16 task states could reliably decode the current state of the task from the OFC, and from that area only (Figure 5C), despite the fact that there were no rewards or reward-expectations in the task. The fidelity of task representations in the OFC correlated with behavioral performance (Figure 5C), and on error trials the correct state was decoded significantly below chance. Moreover, irrelevant aspects of the task (e.g., information from two trials back) were not decodable in the OFC (Figure 5D), supporting the selectivity of the orbitofrontal representation to only task-relevant information, as is required from a minimal state representation. This was in contrast to other brain areas, for instance, the hippocampus and the dorsolateral prefrontal cortex, that represented some (but not all) task-relevant aspects and some task-irrelevant information [6] (Figure 5D). However, recent evidence suggests that replay of sequences of task states in the hippocampus at rest improves orbitofrontal representations [88].

As befits such a critical brain function, the above findings implicate a variety of brain areas in representation learning: attention networks determine perceptual similarity, which is augmented with previously remembered information (e.g., to determine novelty) in the hippocampus, and segmented into latent causes that are represented as separate states in the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

OFC. The dorsolateral prefrontal cortex, in turn, is implicated in switching between state spaces when the global task representation changes [89]. Even with these clues in place, however, it is not clear what brain areas coordinate the online learning of a task-state representation – the elusive process for which we don't yet know the computational algorithm either (Box 1).

Carving the world at its task-relevant joints

Reinforcement learning – how feedback from the world, and particularly unexpected feedback, is incorporated into future predictions – is fairly well understood in the brain. What is less clear is the fundamental process of representation learning – how we learn to carve streams of ongoing experience into task states that correctly encompass all that is relevant to the task at hand in a minimal representation that generalizes learning as widely as possible [27,28]. RL cannot occur without a state representation, and different representations can render a task extremely simple or exquisitely complex. Moreover, since only actions are given feedback in the form of rewards and punishments, the all-important task representation must be learned without direct feedback, extracted from the overall statistics of the task, the environment, and the agent's performance.

In this Perspective, we focused on three lines of work attempting to elucidate the algorithms and neural substrates of representation learning. We first suggested that selective attention, rather than arising from neural constraints, can in fact be viewed as a mechanism that allows rapid learning: selective attention solves the curse of dimensionality in RL by reducing the dimensionality of task states and focusing only on those dimensions that are causally important for the task at hand. By blurring out irrelevant dimensions, selective attention allows us to generalize over them and employ our learning and decision making processes more efficiently. Thus, selective attention helps overcome what is perhaps the most fundamental constraint: we can afford only a limited amount of experience because our lives are finite, and the passage of time is irreversible.

Indeed, within the non-relenting stream of experiences, no two are exactly alike. Thus, learning from past experience entails *generalization*—using experience from one situation to inform us about a (slightly) different situation. In learning the boundaries of generalization, we implicated a clustering process that assumes that similar experiences belong to the same state (cluster), while also allowing for a growing representation when faced with novel experiences [3,52]. We presented empirical evidence suggesting that this similarity-based clustering process is intertwined with learning, and that the learned representations affect both memory and decision making. Learned attention interacts with the clustering process by affecting similarity such that attention to a dimension, e.g., speed of cars, effectively stretches that perceptual axis so that situations with slightly different speeds may be separated into distinct clusters, whereas ignoring a dimension, e.g., color of cars, will mean that different colors will not be considered dissimilar. This similarity then affects both the organization of memory and how new experience is combined with old knowledge, through learning.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

One conclusion from these studies is that, in some cases, slower learning is better. This is particularly important when the goal of learning is to modify previous knowledge when the environment has changed (e.g., from threatening to safe). Fast learning can easily be achieved by postulating a new state of the task, but this prevents the modification of previously held beliefs, associated with the old task states. Our framework suggests that new states are initialized when we are confronted with information that is not similar to previous experience. This unpredictable information would presumably generate a prediction error. A small prediction error will lead to little learning. However, a very large prediction error will lead to generation of a new state. Thus, to impact an old state with new information, the information must be unexpected, but not too much so. This principle of activating the old state and providing non-confirming information to update the predictions contingent on that state, which we took advantage of in our gradual extinction experiment [56], is also the basis of memory-modification methods in cognitive behavioral therapy for post-traumatic stress disorder [90].

The learned state representations are multimodal, potentially incorporating information from any sensory modality, as well as from memory. We reviewed evidence suggesting that the OFC, a prefrontal brain area that receives widespread sensory, limbic and high order afferents [91], is well-poised to represent the abstract identification of the current state [88]. The idea is that the representation in the OFC is akin to a 'pointer' in computer science – an abstract link to information represented in other brain areas, which identifies the current state. Information that is irrelevant to the current state (e.g., the color of oncoming cars) would not affect the orbitofrontal representation. In contrast, information that, if different, would change the current state (e.g., the speed of the closest car), would presumably be decodable in the OFC [92]. Importantly, according to our theory, this is not necessarily because the OFC represents this information *per se*, but because different settings of this variable lead to different states. This hypothesis, different from the dominant view that the OFC represents expected (reward) outcomes [e.g., 93–96], suggests that expected reward will be decodable in the OFC only to the extent that it is part of the state representation (which, in fact, it often is).

In sum, we suggest a dynamic interplay between RL in the basal ganglia, adaptive attention processes in the frontoparietal attention control network, and memory processes that reflect the learned structure of the environment and shape orbitofrontal state representations. In this framework, selective attention is not a limitation of the neural learning system, but an adaptive mechanism that allows rapid learning. Memory is similarly seen as an active process that does not simply mirror the external environment, but rather reflects inference regarding causal relationships in the environment.

Research on representation learning is still in its infancy both in neuroscience, and in machine learning, where a shift is underway from solving specific problems (like playing Go or designing a self-driving car) to designing general artificial intelligence that can adaptively learn to represent and solve new tasks. As such, many critical questions are yet unanswered (Box 1). The findings we have discussed suggest that representation learning involves the dynamic interplay of cognitive functions that have traditionally been studied separately from each other. Understanding how information flows between these systems will help explain

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the amazing adaptive capabilities of humans that go orders of magnitude above and beyond simplified laboratory tasks. In any case, understanding representation learning – this computationally daunting task that our brain so marvelously excels at – will be fundamental to any complete theory of learning in the brain.

Acknowledgements:

I am grateful to my lab members, past and present, for their creative, methodical and incredibly revealing work on representation learning in the brain. I would like to thank Angela Langdon, Nina Rouhani, and Jean Zarate for helpful comments on a previous draft. This work was funded by grant W911NF-14-1-0101 from the Army Research Office and grant R01DA042065 from the National Institute on Drug Abuse.

References

1. Niv Y, Daniel R, Geana A, Gershman SJ, Leong YC, Radulescu A, & Wilson RC (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. Journal of Neuroscience, 35(21), 8145–8157. [PubMed: 26019331]
2. Leong YC, Radulescu A, Daniel R, DeWoskin V, & Niv Y (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. Neuron, 93(2), 451–463. [PubMed: 28103483]
3. Gershman SJ, Blei DM, & Niv Y (2010). Context, learning, and extinction. Psychological Review, 117(1), 197. [PubMed: 20063968]
4. Gershman SJ, Norman KA, & Niv Y (2015). Discovering latent causes in reinforcement learning. Current Opinion in Behavioral Sciences, 5, 43–50.
5. Wilson RC, Takahashi YK, Schoenbaum G, & Niv Y (2014). Orbitofrontal cortex as a cognitive map of task space. Neuron, 81(2), 267–279. [PubMed: 24462094]
6. Schuck NW, Cai MB, Wilson RC, & Niv Y (2016). Human orbitofrontal cortex represents a cognitive map of state space. Neuron, 91(6), 1402–1412. [PubMed: 27657452]
7. Sutton RS, & Barto AG (2018). Reinforcement learning: An introduction. MIT press.
8. Kaelbling LP, Littman ML, & Moore AW (1996). Reinforcement learning: A survey. Journal of artificial intelligence research, 4, 237–285.
9. Daw ND, & Tobler PN (2014). Value learning through reinforcement: the basics of dopamine and reinforcement learning In: Glimcher PW & Fehr E, eds: Neuroeconomics. Academic Press 283–298.
10. Daw ND, & O'Doherty JP (2014). Multiple systems for value learning In: Glimcher PW & Fehr E, eds: Neuroeconomics. Academic Press 393–410.
11. Niv Y, & Langdon A (2016). Reinforcement learning with Marr. Current Opinion in Behavioral Sciences, 11, 67–73. [PubMed: 27408906]
12. Watkins CJ, & Dayan P (1992). Q-learning. Machine learning, 8(3–4), 279–292.
13. Friedrich J, & Lengyel M (2016). Goal-directed decision making with spiking neurons. Journal of Neuroscience, 36(5), 1529–1546. [PubMed: 26843636]
14. Daw ND, Niv Y, & Dayan P (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature neuroscience, 8(12), 1704–1711. [PubMed: 16286932]
15. Keramati M, Smittenaar P, Dolan RJ, & Dayan P (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. Proceedings of the National Academy of Sciences, 113(45), 12868–12873.
16. Barto AG (1995). Adaptive critics and the basal ganglia In: Houk JC and Davis JL and Beiser DG, Eds. Models of information processing in the basal ganglia, 215–232.
17. Montague PR, Dayan P, & Sejnowski TJ (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. Journal of neuroscience, 16(5), 1936–1947. [PubMed: 8774460]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

18. Yin HH, Knowlton BJ, & Balleine BW (2006). Inactivation of dorsolateral striatum enhances sensitivity to changes in the action–outcome contingency in instrumental conditioning. Behavioural brain research, 166(2), 189–196. [PubMed: 16153716]
19. Miller KJ, Botvinick MM, & Brody CD (2017). Dorsal hippocampus contributes to model-based planning. Nature neuroscience, 20(9), 1269. [PubMed: 28758995]
20. Vikbladh OM, Meager MR, King J, Blackmon K, Devinsky O, Shohamy D, … & Daw ND (2019). Hippocampal Contributions to Model-Based Planning and Spatial Memory. Neuron, 102(3), 683–693. [PubMed: 30871859]
21. McDannald MA, Lucantonio F, Burke KA, Niv Y, & Schoenbaum G (2011). Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. Journal of Neuroscience, 31(7), 2700–2705. [PubMed: 21325538]
22. Boorman ED, Rajendran VG, O'Reilly JX, & Behrens TE (2016). Two anatomically and computationally distinct learning signals predict changes to stimulus-outcome associations in hippocampus. Neuron, 89(6), 1343–1354. [PubMed: 26948895]
23. Kempadoo KA, Mosharov EV, Choi SJ, Sulzer D, & Kandel ER (2016). Dopamine release from the locus coeruleus to the dorsal hippocampus promotes spatial learning and memory. Proceedings of the National Academy of Sciences, 113(51), 14835–14840.
24. Rouhani N, Norman KA, & Niv Y (2018). Dissociable effects of surprising rewards on learning and memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 44(9), 1430–1443.
25. Langdon AJ, Sharpe MJ, Schoenbaum G, & Niv Y (2018). Model-based predictions for dopamine. Current Opinion in Neurobiology, 49, 1–7. [PubMed: 29096115]
26. Ponsen M, Taylor ME, Tuyls K (2010) Abstraction and Generalization in Reinforcement Learning: A Summary and Framework In: Taylor ME, Tuyls K (eds) Adaptive and Learning Agents. ALA 2009. Lecture Notes in Computer Science, vol 5924 Springer, Berlin, Heidelberg DOI: 10.1007/978-3-642-11814-2_1
27. Canas F, & Jones M (2010). Attention and reinforcement learning: constructing representations from indirect feedback. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 32, No. 32). Permalink: https://escholarship.org/uc/item/1w83t8ct
28. Jones M, & Canas F (2010). Integrating reinforcement learning with models of representation learning. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 32, No. 32). Permalink: https://escholarship.org/uc/item/88×6f84q
29. Bellman R (1957). Dynamic Programming (Princeton University Press)
30. Sutton RS (1996). Generalization in reinforcement learning: Successful examples using sparse coarse coding. In Advances in neural information processing systems (pp. 1038–1044).
31. Tesauro G (1994). TD-Gammon, a self-teaching backgammon program, achieves master-level play. Neural computation, 6(2), 215–219
32. Ludvig EA, Sutton RS, & Kehoe EJ (2012). Evaluating the TD model of classical conditioning. Learning & behavior, 40(3), 305–319. [PubMed: 22927003]
33. McCallum RA (1996). Hidden state and reinforcement learning with instance-based state identification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 26(3), 464–473.
34. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, … & Chen Y (2017). Mastering the game of go without human knowledge. Nature, 550(7676), 354. [PubMed: 29052630]
35. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, … & Petersen S (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529. [PubMed: 25719670]
36. Lake BM, Ullman TD, Tenenbaum JB, & Gershman SJ (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40.
37. Wang JX, Kurth-Nelson Z, Tirumala D, Soyer H, Leibo JZ, Munos R, Blundell C, Kumaran D, & Botvinick M (2016). Learning to reinforcement learn, arXiv:1611.05763

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

38. Bramley NR, Dayan P, Griffiths TL, & Lagnado DA (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. Psychological Review, 124(3), 301. [PubMed: 28240922]
39. Griffiths TL, Chater N, Kemp C, Perfors A, & Tenenbaum JB (2010). Probabilistic models of cognition: Exploring representations and inductive biases. Trends in cognitive sciences, 14(8), 357–364. [PubMed: 20576465]
40. Dias R, Robbins TW, & Roberts AC (1996). Primate analogue of the Wisconsin Card Sorting Test: effects of excitotoxic lesions of the prefrontal cortex in the marmoset. Behavioral neuroscience, 110(5), 872. [PubMed: 8918991]
41. Frank MJ, Seeberger LC, & O'reilly RC (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science, 306 (5703), 1940–1943. [PubMed: 15528409]
42. Daw ND, O'Doherty JP, Dayan P, Seymour B, & Dolan RJ (2006). Cortical substrates for exploratory decisions in humans. Nature, 441(7095), 876. [PubMed: 16778890]
43. Milner B (1963). Effects of different brain lesions on card sorting. Archives of Neurology, 9, 100–110.
44. Kruschke JK (1992). ALCOVE: an exemplar-based connectionist model of category learning. Psychological Review, 99(1), 22. [PubMed: 1546117]
45. Petersen SE, and Posner MI (2012). The attention system of the human brain: 20 years after. Annual Reviews in Neuroscience, 35, 73–89.
46. Corbetta M, & Shulman GL (2002). Control of goal-directed and stimulus-driven attention in the brain. Nature reviews neuroscience, 3(3), 201. [PubMed: 11994752]
47. Kruschke JK (2005). Learning involves attention In: Connectionist models in cognitive psychology, Editor: Houghton G. 113–140. Psychology Press.
48. Kruschke JK (2001). Toward a unified model of attention in associative learning. Journal of mathematical psychology, 45(6), 812–863.
49. McCallum RA (1995). Instance-based utile distinctions for reinforcement learning with hidden state In Machine Learning Proceedings 1995 (pp. 387–395). Morgan Kaufmann.
50. Collins AG & Frank MJ (2013). Cognitive control over learning: creating, clustering, and generalizing task-set structure, Psychological Review, 120(1):190. [PubMed: 23356780]
51. Langdon AJ, Song M & Niv Y (in press) Uncovering the 'state': tracing the hidden representations that structure learning and decision-making, Behavioral Processes
52. Love BC, Medin DL, & Gureckis TM (2004). SUSTAIN: a network model of category learning. Psychological Review, 111(2), 309. [PubMed: 15065912]
53. Rescorla RA (2004). Spontaneous recovery. Learning & Memory, 11(5), 501–509. [PubMed: 15466300]
54. Bouton ME (2004). Context and behavioral processes in extinction. Learning & Memory, 11(5), 485–494. [PubMed: 15466298]
55. Rescorla RA, & Heth CD (1975). Reinstatement of fear to an extinguished conditioned stimulus. Journal of Experimental Psychology: Animal Behavior Processes, 1(1), 88. [PubMed: 1151290]
56. Gershman SJ, Jones CE, Norman KA, Monfils MH, & Niv Y (2013). Gradual extinction prevents the return of fear: implications for the discovery of state. Frontiers in behavioral neuroscience, 7, 164. [PubMed: 24302899]
57. Gershman SJ, & Hartley CA (2015). Individual differences in learning predict the return of fear. Learning & Behavior, 43(3), 243–250. [PubMed: 26100524]
58. Gershman SJ, & Niv Y (2013). Perceptual estimation obeys Occam's razor. Frontiers in Psychology, 4, 623. [PubMed: 24137136]
59. Preminger S, Blumenfeld B, Sagi D, & Tsodyks M (2009). Mapping dynamic memories of gradually changing objects. Proceedings of the National Academy of Sciences, 106(13), 5371–5376.
60. Gershman SJ, Radulescu A, Norman KA, & Niv Y (2014). Statistical computations underlying the dynamics of memory updating. PLoS computational biology, 10(11), e1003939. [PubMed: 25375816]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

61. Gershman SJ, Monfils MH, Norman KA, & Niv Y (2017). The computational nature of memory modification. eLife, 6, e23763. [PubMed: 28294944]
62. Ji J, & Maren S (2007). Hippocampal involvement in contextual modulation of fear extinction. Hippocampus, 17(9), 749–758. [PubMed: 17604353]
63. Honey RC, & Good M (1993). Selective hippocampal lesions abolish the contextual specificity of latent inhibition and conditioning. Behavioral neuroscience, 107(1), 23. [PubMed: 8447955]
64. Yap CS, & Richardson R (2007). Extinction in the developing rat: an examination of renewal effects. Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology, 49(6), 565–575.
65. Yap CS, & Richardson R (2005). Latent inhibition in the developing rat: An examination of context-specific effects. Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology, 47(1), 55–65.
66. Knight RT (1996). Contribution of human hippocampal region to novelty detection. Nature, 383(6597), 256. [PubMed: 8805701]
67. Kumaran D, & Maguire EA (2007). Which computational mechanisms operate in the hippocampus during novelty detection? Hippocampus, 17(9), 735–748. [PubMed: 17598148]
68. Mednick SA, & Lehtinen LE (1957). Stimulus generalization as a function of age in children. Journal of Experimental Psychology, 53(3), 180. [PubMed: 13416480]
69. Droit-Volet S, Clément A, & Wearden J (2001). Temporal generalization in 3-to 8- year-old children. Journal of Experimental Child Psychology, 80(3), 271–288. [PubMed: 11583526]
70. Wikenheiser AM, & Schoenbaum G (2016). Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. Nature Reviews Neuroscience, 17(8), 513. [PubMed: 27256552]
71. Stalnaker TA, Cooch NK, & Schoenbaum G (2015). What the orbitofrontal cortex does not do. Nature neuroscience, 18(5), 620. [PubMed: 25919962]
72. Izquierdo A, Suda RK, & Murray EA (2004). Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. Journal of Neuroscience, 24(34), 7540–7548. [PubMed: 15329401]
73. Chudasama Y, & Robbins TW (2003). Dissociable contributions of the orbitofrontal and infralimbic cortex to pavlovian autoshaping and discrimination reversal learning: further evidence for the functional heterogeneity of the rodent frontal cortex. Journal of Neuroscience, 23(25), 8771–8780. [PubMed: 14507977]
74. Walton ME, Behrens TE, Buckley MJ, Rudebeck PH, & Rushworth MF (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. Neuron, 65(6), 927–939. [PubMed: 20346766]
75. Tsuchida A, Doll BB, & Fellows LK (2010). Beyond reversal: a critical role for human orbitofrontal cortex in flexible learning from probabilistic feedback. Journal of Neuroscience, 30(50), 16868–16875. [PubMed: 21159958]
76. Lak A, Costa GM, Romberg E, Koulakov AA, Mainen ZF, & Kepecs A (2014). Orbitofrontal cortex is required for optimal waiting based on decision confidence. Neuron, 84(1), 190–201. [PubMed: 25242219]
77. Takahashi YK, Roesch MR, Wilson RC, Toreson K, O'donnell P, Niv Y, & Schoenbaum G (2011). Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. Nature neuroscience, 14(12), 1590. [PubMed: 22037501]
78. Blanchard TC, Hayden BY, & Bromberg-Martin ES (2015). Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. Neuron, 85(3), 602–614. [PubMed: 25619657]
79. Stalnaker TA, Cooch NK, McDannald MA, Liu TL, Wied H, & Schoenbaum G (2014). Orbitofrontal neurons infer the value and identity of predicted outcomes. Nature communications, 5, 3926.
80. Farovik A, Place RJ, McKenzie S, Porter B, Munro CE, & Eichenbaum H (2015). Orbitofrontal cortex encodes memories within value-based schemas and represents contexts that guide memory retrieval. Journal of Neuroscience, 35(21), 8333–8344. [PubMed: 26019346]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

81. Zhou J, Gardner MP, Stalnaker TA, Ramus SJ, Wikenheiser AM, Niv Y, & Schoenbaum G (2019). Rat Orbitofrontal Ensemble Activity Contains Multiplexed but Dissociable Representations of Value and Task Structure in an Odor Sequence Task. Current Biology, 29(6), 897–907. [PubMed: 30827919]
82. Howard JD, Gottfried JA, Tobler PN, & Kahnt T (2015). Identity-specific coding of future rewards in the human orbitofrontal cortex. Proceedings of the National Academy of Sciences, 112(16), 5195–5200.
83. Chan SC, Niv Y, & Norman KA (2016). A probability distribution over latent causes, in the orbitofrontal cortex. Journal of Neuroscience, 36(30), 7817–7828. [PubMed: 27466328]
84. Hampton AN, Bossaerts P, & O'Doherty JP (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. Journal of Neuroscience, 26(32), 8360–8367. [PubMed: 16899731]
85. Takahashi YK, Langdon AJ, Niv Y, & Schoenbaum G (2016). Temporal specificity of reward prediction errors signaled by putative dopamine neurons in rat VTA depends on ventral striatum. Neuron, 91(1), 182–193. [PubMed: 27292535]
86. Bradfield LA, Dezfouli A, van Holstein M, Chieng B, & Balleine BW (2015). Medial orbitofrontal cortex mediates outcome retrieval in partially observable task situations. Neuron, 88(6), 1268–1280. [PubMed: 26627312]
87. Takahashi YK, Stalnaker TA, Roesch MR, & Schoenbaum G (2017). Effects of inference on dopaminergic prediction errors depend on orbitofrontal processing. Behavioral Neuroscience, 131(2), 127. [PubMed: 28301188]
88. Schuck NW, & Niv Y (2019). Sequential replay of nonspatial task states in the human hippocampus. Science, 364(6447), eaaw5181.
89. Sharpe MJ, Stalnaker T, Schuck NW, Killcross S, Schoenbaum G, & Niv Y (2019). An integrated model of action selection: distinct modes of cortical control of striatal decision making. Annual review of psychology, 70, 53–76.
90. Foa EB, & Kozak MJ (1986). Emotional processing of fear: exposure to corrective information. Psychological bulletin, 99(1), 20. [PubMed: 2871574]
91. Wallis JD (2007). Orbitofrontal cortex and its contribution to decision-making. Annu. Rev. Neurosci, 30, 31–56. [PubMed: 17417936]
92. Schoenbaum G, Setlow B, & Ramus SJ (2003). A systems approach to orbitofrontal cortex function: recordings in rat orbitofrontal cortex reveal interactions with different learning systems. Behavioural brain research, 146(1–2), 19–29. [PubMed: 14643456]
93. Padoa-Schioppa C, & Assad JA (2006). Neurons in the orbitofrontal cortex encode economic value. Nature, 441(7090), 223. [PubMed: 16633341]
94. Padoa-Schioppa C (2011). Neurobiology of economic choice: a good-based model. Annual review of neuroscience, 34, 333–359.
95. Plassmann H, O'Doherty J, & Rangel A (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. Journal of neuroscience, 27(37), 9984–9988. [PubMed: 17855612]
96. McNamee D, Rangel A, & O'doherty JP (2013). Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex. Nature Neuroscience, 16(4), 479. [PubMed: 23416449]
97. Nosofsky RM (1986). Attention, similarity, and the identification–categorization relationship. Journal of experimental psychology: General, 115(1), 39. [PubMed: 2937873]
98. Summerfield C, & De Lange FP (2014). Expectation in perceptual decision making: neural and computational mechanisms. Nature Reviews Neuroscience, 15(11), 745. [PubMed: 25315388]
99. Colgin LL, Moser EI, & Moser MB (2008). Understanding memory through hippocampal remapping. Trends in neurosciences, 31(9), 469–477. [PubMed: 18687478]
100. Leutgeb JK, Leutgeb S, Treves A, Meyer R, Barnes CA, McNaughton BL, … & Moser EI (2005). Progressive transformation of hippocampal neuronal representations in "morphed" environments. Neuron, 48(2), 345–358. [PubMed: 16242413]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Box 1.

Open questions: representation learning in real time, in the brain

A major outstanding question regards how trial by trial feedback shapes attention (which, in turn, shapes task representations). Our findings show that the interaction between attention and learning is bidirectional, with high-value features attracting attention and attention determining how value is accrued to different features [2]. Still, they do not explain how attention is determined trial-by-trial, that is, how feedback (e.g., rewards and prediction errors) is used to determine whether to switch the focus of attention on the next instance [27–28,48].

Computationally, it is not known what statistics of the task (and our performance on it) convey that we are focusing on the wrong task dimensions [97]? What evidence is indicative of too narrow a focus of attention that must be widened, and vice versa? It is likely that this type of evidence transcends a single trial. For instance, one way to define the quality of a representation is by the entropy of motivationally important outcomes predicated on that representation, with lower entropy (i.e., more deterministic predictions) being preferable. For example, the representation in Figure 1B is less effective than that in Figure 1C in predicting our chances of reaching the other side of the street safely. However, this, and similar hypotheses, require testing.

Neurally, it is not yet known whether selective attention operates directly to shape the funneling-in cortical afferents to the striatum (so that irrelevant, unattended, input dimensions do not contribute to striatal RL as the striatum does not “know” that these dimensions existed in stimuli), computations within the striatum (e.g., through weighting the contribution of different input dimensions to the valuation of the current state), or cortical representations (e.g., in the OFC) which, in turn, affect what the striatum represents and learns about. In perception, expectations and attention affect neural responses to stimuli both before (with elevated activity for predicted stimuli) and after stimulus onset (with predicted stimuli eliciting less activation than surprising ones due to ‘expectation suppression’) [98]. These neural effects have been formalized within the computational framework of predictive coding, as effects on initial starting points and drift rates in Bayesian evidence accumulation models [98]. While we have shown that attention to input dimensions influences both valuation and updating in RL [2], an understanding of how this manifests neurally, and indeed a detailed comparison of computational models of this influence, await future work.

Other open questions include what brain areas are involved in learning new representations (rather than representing known task states), and how arbitrary ‘pointers’ to task states that can differ radically from task to task are implemented in the OFC. Better insight into the nature of state representations in the OFC will hopefully help explain how the OFC can switch rapidly and dynamically between representations of different tasks, and, indeed, how completely new task representations can be generated in the OFC in moments, as people are explicitly instructed about the structure of a new task (e.g., as was done in [6]). Ideas about how the hippocampus represents spatial maps that differ from environment to environment [99–100] may provide important clues regarding

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the representation of abstract cognitive maps of tasks. Moreover, understanding how unexpected feedback shapes learning of task representations may shed light on, for example, the role of dopamine in prefrontal cortices and in the hippocampus.

An important question that the surveyed research has not addressed is how does the general structure of inputs (some of which are not task-relevant), which is learned through unsupervised learning, interact with learning of task-relevant representations. For instance, in the Dimensions Task, all features are present on every trial, but this is not the case when crossing the street. An important hint that the barking dog is irrelevant for the latter task is probably the fact that street crossing is seldom accompanied by a barking dog.

Finally, we do not yet know what constraints our brain has adapted to and been able to take advantage of: what are the statistics of natural tasks? Can the brain safely assume that even in highly multidimensional environments only a few dimensions are relevant to any given task? Presumably, decision-making systems evolved to be tailored to the set of tasks that we are routinely faced with. Similar to the breakthroughs in understanding vision that followed the quantification of statistics of natural scenes, a clear description of the statistics of natural tasks might revolutionize our understanding of the neural basis of high-level learning and decision making.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

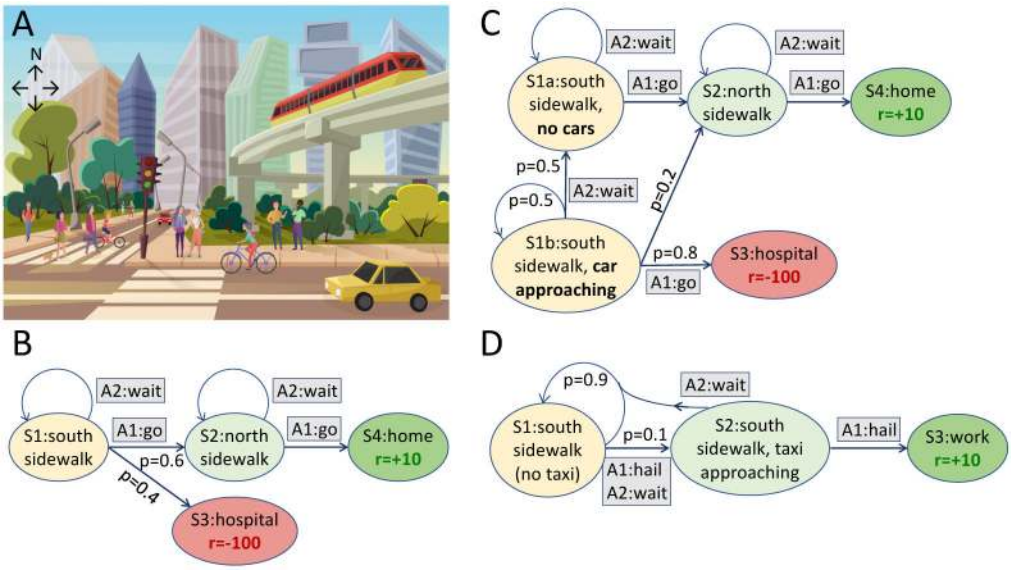

Figure 1. Task representations.

A) The setting: a street corner. You are on the south side just outside the picture. **B)** State diagram of the task of "going home" from your egocentric point of view. States are in circles, actions in gray rectangles, arrows denote state transitions, reward outcomes in color within the respective state (r=0 if not stated). At the first state (S1, yellow: south sidewalk), choosing A1 (go) will result in a transition to S2 (light green; north sidewalk) with probability 0.6, and a transition to S3 (red; hospital, due to being run over) with probability 0.4. Since hospitalization is accompanied by a very aversive outcome (r=−100) that overwhelms the appetitive outcome of making it home safely (S4, green, r=10), the optimal policy is to wait at S1 indefinitely. **C)** An alternative state representation for the task of "going home" that separates S1, the state of standing on the south sidewalk, to two states: S1a where no car is in sight and crossing the street is perfectly safe, and S1b where a car is approaching and crossing the street is dangerous (only p=0.2 for the transition to the north sidewalk when choosing A1, and p=0.8 for the transition to the hospital). Waiting is the optimal policy in S1b, and going is optimal at S1a. Eventually, you will get home safely. **D)** Representation of an alternative task in the same exact setting: hailing a taxi to go to work. Transitions in S1 occur regardless of your chosen action (to hail or to wait), however, in S2 the action of hailing a taxi will get you to work whereas waiting will bring you back to S1.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

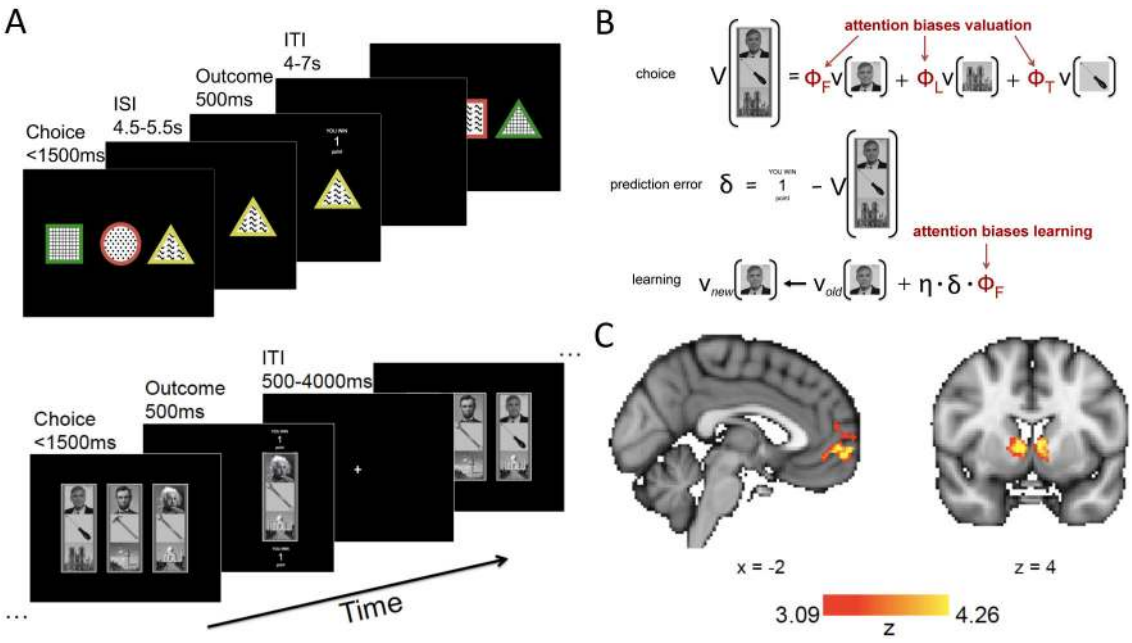

Figure 2. The Dimensions Task.

A) Two variants of the Dimensions Task, one with overlaid color, shape and texture dimensions (top), and another with rectangle stimuli comprising of face, tool and scene dimensions (bottom) that allow direct measurement of attention via eye-tracking and multivariate pattern analysis from visual areas responsive to faces, tools and scenes. **B)** Attention weighted reinforcement learning. A value is learned for each of the 9 features. The values are combined, weighted by attention to each of the face, tool and scene dimensions, to determine the total value (expected reward) of a rectangle stimulus. Once the outcome is observed, a prediction error is calculated as the difference between the outcome to the expected reward (value). This prediction error is used to update the values of each of the chosen features, weighed by attention to each dimension. ϕ, attention weight; δ, prediction error; η, learning rate or step size parameter. **C)** than Neural activations corresponding to areas that correlate with values (left) and prediction errors (right) from the attention-weighted algorithm significantly better than from a similar algorithm with uniform (1/3) weights for each of the dimensions. Figure adapted from [1–2].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

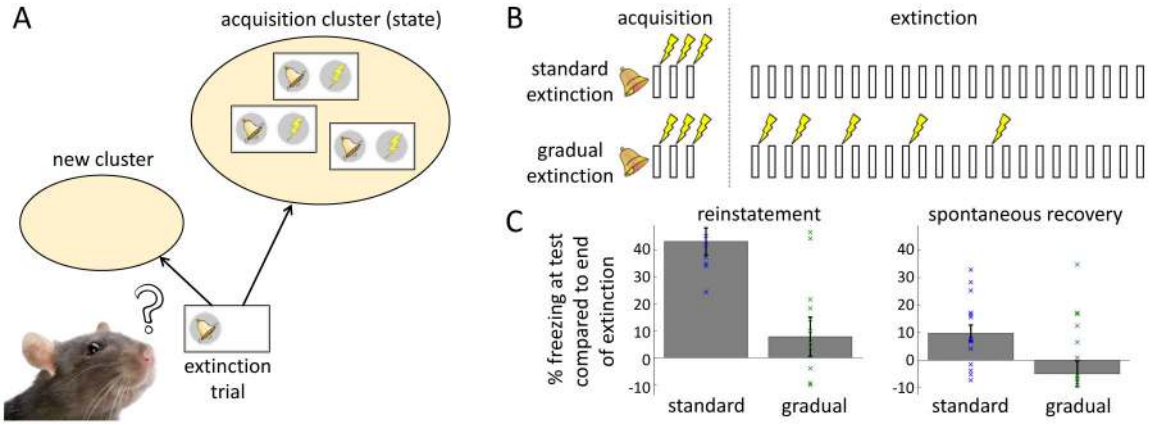

Figure 3. Latent-cause inference in representation learning suggests a mechanism for altering fear memories.

A) When confronted with the first extinction trial, the animal determines, based on the (dis)similarity of this trial to previous experiences (acquisition trials) whether to attribute it to the acquisition cluster or a new cluster. Clusters correspond to states, and are determined based on inferred latent causes. In the inference process, cluster assignment is probabilistic such that an event can be partially assigned to more than one cluster. **B)** A gradual extinction protocol aimed at making extinction more similar to acquisition, and thus coaxing the animal to assign all trials to a single cluster. Top: standard extinction, with no shocks in the extinction phase. Bottom: gradual extinction where shocks taper off gradually in the extinction phase. **C)** In two experiments, rats showed significant return of fear after reinstatement with a reminder shock (left), or due to spontaneous recovery (right) following standard extinction (freezing at test compared to last 4 trials of extinction), whereas after gradual extinction rats show no increase of fear at test. Figure adapted from [56].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

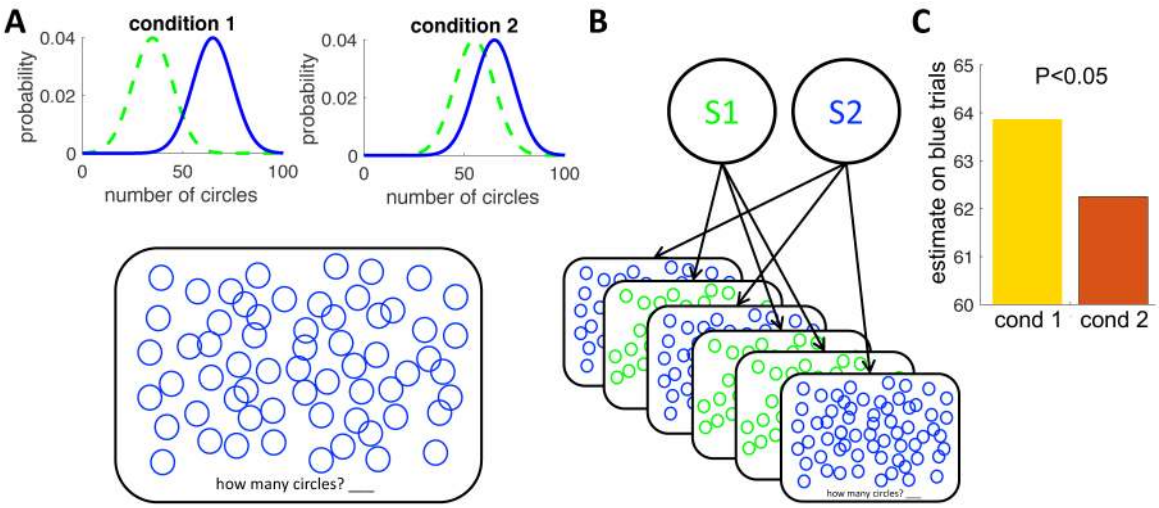

Figure 4. Humans spontaneously use similarity to infer the latent structure of task.

A) The Circles Task [58]. Human participants were asked to key in a double-digit guess for the number of circles on the screen (bottom). In condition 1, the number of circles was drawn either from a Gaussian of mean 65 or from a Gaussian of mean 35 (top left). In condition 2 means were 65 and 55 (top right). Conditions occurred in blocks, with each Gaussian associated with a different color of the circles. Blocks of the two conditions were randomly intermixed within participant (8 blocks of each, 20 stimuli per block). Each block involved two different colors. We use only blue and green to illustrate that the mean-65 trials were identical in both conditions. **B)** Participants' guesses on the mean-65 trials in the two conditions suggested that they spontaneously inferred from the whether the task involved two latent causes (states), each generating stimuli of one color (depicted) or rather there was only one latent cause generating all trials in a block. **C)** Because of the greater overlap between the distributions in condition 2, which resulted in stimuli that were more similar across colors, in that case participants tended to infer a single latent cause, thus ignoring the color of stimuli and guessing the number of circles on blue trials as closer to the global mean of 60. In condition 1, stimuli were sufficiently different to warrant two latent causes, such that learning about blue circles was segregated from learning about green circles, and the estimate on blue trials was closer to their true mean of 65.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

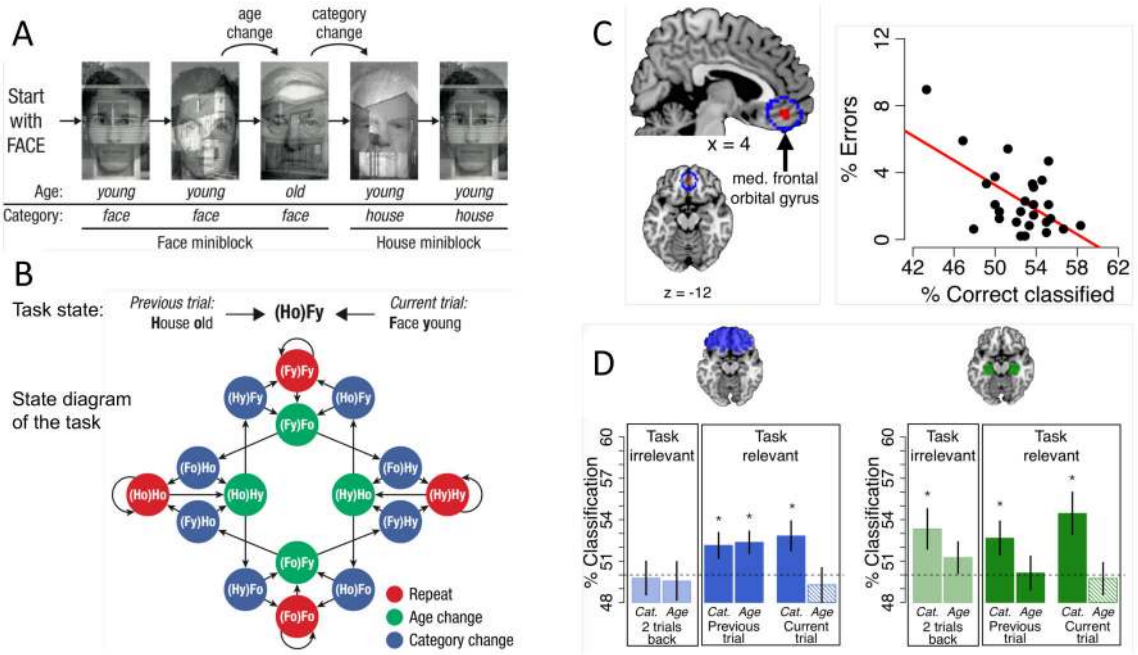

Figure 5. The orbitofrontal cortex represents the current state of the task.

A) An age-judgement task with hidden states. Participants must judge the age (young/old) of one category (e.g., faces) until the age in that category changes. The subsequent trial then starts a miniblock of judging the alternate category (e.g., houses) until the age in that category changes, and so forth. **B)** These instructions create a 16-state task where each state includes the current category to be judged, the age of the current stimulus in that category, the age in the previous trial (for comparison with current age) and the category in the previous trial (as age comparison is not needed in the beginning of a miniblock). All state components are unobservable except for current age. Each state transitions to one of two other states, with equal probability. **C)** The orbitofrontal cortex (left) was the only brain area from which all unobservable components could be classified, and classification accuracy there (in an anatomically defined region of interest including the whole OFC) correlated with lower behavioral error rates (right). **D)** Only unobservable, task relevant, features were decodable in the OFC (left) in contrast to the hippocampus where only category was decodable, for several trials back (right). Error bars: SEM; dashed horizontal: chance baseline; * $p<0.05$ compared to chance, one-tailed. Figure adapted from [6] and [87].