How working memory and reinforcement learning are intertwined: a cognitive, neural, and computational perspective.

Abstract

Reinforcement learning and working memory are two core processes of human cognition, and are often considered cognitively, neuroscientifically, and algorithmically distinct. Here, we show that the brain networks that support them actually overlap significantly, and that they are less distinct cognitive processes than often assumed. We review literature demonstrating the benefits of considering each process to explain properties of the other, and highlight recent work investigating their more complex interactions. We discuss how future research in both computational and cognitive sciences can benefit from one another, suggesting that a key missing piece for artificial agents to learn to behave with more human-like efficiency is taking working memory's role in learning seriously. This review highlights the risks of neglecting the interplay between different processes when studying human behavior (in particular when considering individual differences). We emphasize the importance of investigating these dynamics in order to build a comprehensive understanding of human cognition.

Introduction

Reinforcement learning (RL) and working memory (WM) are two core processes of human cognition. RL broadly refers to a set of behavioral, neuroscientific, and computational processes in which an agent learns through trial and error with the goal of maximizing reward (Eckstein et al., 2021; Sutton & Barto, 1998). WM refers to an information-limited process used to hold representations in the mind temporarily for use in thought and action (Cowan, 2017; Oberauer et al., 2018). They are essential in a range of daily activities that require intelligent, flexible behavior. Deficits in RL and WM are related to cognitive decline and often observed in mental disorders such as schizophrenia and depression. While there is a rich body of literature investigating each process separately, the aim of this review is to discuss the relationship

between them. Specifically, we review literature explaining the neural, behavioral, and computational interplay between the two systems, and discuss the importance of paying attention to one process when investigating the other.

In the first section, "Defining RL and WM," we will describe each process independently, in terms of the behaviors they support, the neural representations underlying them, and the computational models developed to characterize them. In the second section, "The interplay of RL and WM," we will show that these two processes are related neurally and behaviorally, and that both processes can be better understood when considering how the other affects it. In the third section, "The importance of investigating inter-process dynamics," we discuss how only considering one process can misrepresent data, and thus lead to incorrect conclusions. Finally, we will discuss their interactions with other processes (in "Interactions with other processes"), and the insights about cognition and neuroscience that can be gained by investigating recent efforts in the field of artificial intelligence to make agents' behavior better match humans' ability to learn, generalize, and make flexible decisions (in "Computational insights"). The goal of this paper is to review the research investigating the relationship between two seminal processes, and highlight how investigating the richness of their interplay is important to developing veridical computational and neural understandings of behavior across a variety of contexts.

Defining RL and WM

Before discussing the relationship between these two processes, we review each in isolation. We find it of particular importance to discuss how three overlapping, but distinct, subfields define each process: psychology and cognitive sciences, focusing on behavior; neuroscience, focusing on brain networks; and computational fields such as artificial intelligence, focusing on algorithms (Eckstein et al., 2021). We attempt to highlight the risks that come from multiple subfields using the same term (e.g., "RL") with subtly different meanings, and to remove this ambiguity. For example, an RL behavioral task may not be best explained by an RL algorithm alone, and may not only rely on the brain's RL network. More examples of these will be expanded in the next section ("The interplay of RL and WM").

RL behavior, RL brain, RL algorithms

RL in psychology describes animals' ability to learn to make choices to seek rewards and avoid punishments, and has a long history rooted in behaviorism (e.g., Rescorla & Wagner, 1972) that has since been largely expanded to more complex cognitive processes. RL describes the process that allows agents to gradually integrate a past history of reward outcomes into a robust choice policy that supports good decisions.

We define "RL behavior" as learning behavior in sequential decision making tasks that involve appetitive or aversive outcomes. Behavioral RL tasks range in complexity and structure, but always involve the participant learning the value (or a proxy) of an action, state, or series of actions and states through trial and error, using a form of reward or punishment as a teaching signal (e.g., food, points, money, pain). Some common RL tasks are learning which keypress results in a reward for a particular stimulus (stimulus-action association), navigating to a goal state in a grid world or maze, or discovering which of several options results in the highest expected reward (bandit tasks; Figure 1A). All of these tasks are alike in that the goal is to maximize rewards, which are learned incrementally from valenced feedback.

Neurally, RL is widely thought to be supported by dopaminergic signaling in the basal ganglia, particularly in the striatum (e.g., Houk, 1995; Joel et al., 2002; Schultz et al., 1995, 1997; Suri et al., 2001; Sutton & Barto, 1998). A dominant theory of RL in the brain is the reward prediction error theory of dopamine (e.g., Bayer & Glimcher, 2005; Maes et al., 2020; Satoh et al., 2003; Daw & Tobler, 2014; Montague et al., 1996; Niv, 2009; Schultz, 2002; Schultz et al., 1997). In this theory, supported by a broad range of findings, phasic dopaminergic activity encodes the temporal difference *reward prediction error* (*RPE*; a difference between future outcome expectations at different time points; Figure 1B top inset). In other words, the spiking of dopaminergic neurons in basal ganglia is increased when a reward is larger than expected (a positive RPE) and decreased when a reward is smaller than expected (a negative RPE). This RPE-encoding signal facilitates cortico-basal ganglial plasticity (Wickens, 2009) and allows striatal neurons to learn to encode choice values (Samejima et al., 2005), supporting a choice strategy that favors choices that usually lead to better outcomes.

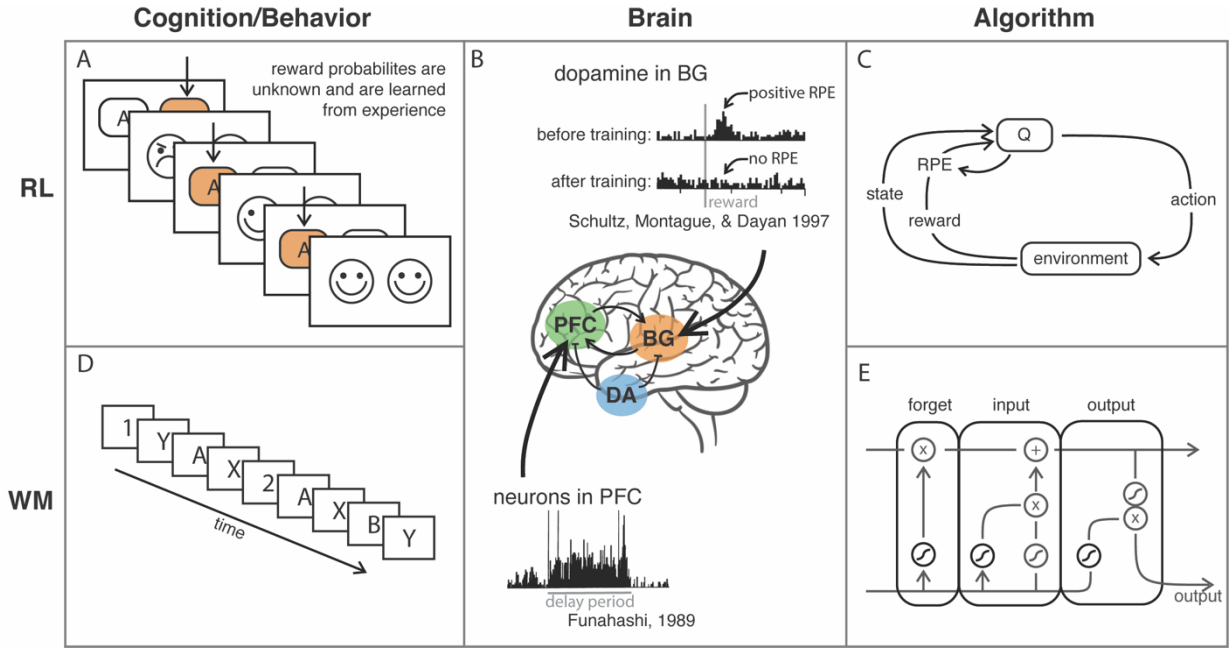

Figure 1. RL (top) and WM (bottom) processes are associated with a broad range of behavioral paradigms (left; A,D), brain areas (middle; B), and algorithms (right; C,E). A. A schematic of a multi-armed bandit problem. The participant must iteratively learn from feedback which option has the highest expected reward. B. RL and WM rely on overlapping brain networks, both modulated by dopamine (DA), although WM is largely described to be a prefrontal cortex (PFC) associated area, and RL a basal ganglia (BG) associated area (black arrows). top inset: dopaminergic activity reflects reward prediction error (RPE). bottom inset: elevated delay-period activity in PFC while maintaining information in WM. C. A schematic of a general RL agent that learns the value of different state action pairs (the Q-value) iteratively using RPE. D. A schematic of the 1-2-AX task, in which the participant must selectively maintain letters based on context (numbers). E. A schematic of the LSTM model, in which WM representations can be independently forgotten, inputted, and outputted.

RL is also a broad area of machine learning and *artificial intelligence* (*AI*). RL AI represents a family of learning algorithms primarily used in sequential/multistep problems (called *Markov Decision Processes* or *MDPs)*, where the current state of the world is fully informative of which action an agent should take (Sutton & Barto, 1998). RL artificial agents share the property that

they rely on algorithms whose objective is to learn to make choices that maximize future cumulative rewards. This can be achieved with broadly different algorithms, for example by trying to estimate the value of choices by incrementally updating it when rewards are observed, in proportion to the reward prediction error, or by using information about the environment to effortfully compute the expected future value of a choice (Figure 1C illustrates the updating the expected value of a certain action and state, referred to as the Q-value, based on the reward received). A lot of RL research in machine learning has no bearing to cognitive psychology and neuroscience (e.g., autonomous navigation of stratospheric balloons, Bellemare et al., 2020); however, an important subset of RL algorithms have been extremely successful at describing both RL behavior and RL in the brain (Montague et al., 1996; Schultz et al., 1997).

WM behavior, WM brain, WM algorithms

Similar to the RL process, the WM process should be explicitly defined in different subfields. WM broadly refers to an information-limited process used to hold a small amount of information in mind for a small amount of time, when that information is no longer perceptually available, for use in thought and action (a classic example is the memorization of a phone number). When we refer to WM, we refer to the process that allows agents to both store and manipulate information, which supports abstract, goal-directed behavior. Thus, we do not only consider WM as a passive storage unit, but also closely related to executive function.

WM behavioral tasks (for example Figure 1D) all involve participants holding some number of representations in mind over a delay; participants are later tested on their retention either directly (enter the phone number) or via a manipulation (enter the phone number backwards). One canonical effect in the WM literature regardless of modality is the decreased accuracy and increased response with increasing number of memoranda (i.e., the *"set size" effect*; e.g., Sternberg, 1966; Luck & Vogel, 1997). This effect demonstrates one of the defining features of WM: its information-limited capacity. Despite its limited capacity, people are able to selectively maintain information more behaviorally relevant (Bays & Husain, 2008; Braver & Cohen, 2000), demonstrating the ability of WM to only "gate in" desired information. Another canonical effect is the decreased accuracy with increasing WM delay times and/or distractors, demonstrating the fragility of WM representations (e.g., Brown, 1958; Peterson and Peterson, 1959). These

behavioral characteristics of WM are in contrast to those of other, longer-term memory mechanisms, which are not capacity limited and do not require active maintenance to later recall information. We recommend the review article by Oberauer et al. (2018) for a comprehensive overview of different behavioral benchmarks of WM, across modalities and experimental paradigms.

While "WM brain" is canonically characterized as elevated, persistent neural activity in the prefrontal cortex (Baddeley & Hitch, 1974; Funahashi et al., 1989; Fuster & Alexander, 1971; Figure 1B, bottom inset), recent neuroimaging studies have demonstrated that WM may be represented without elevated, persistent activity (e.g., Murray et al., 2017; Stokes, 2015; Christophel et al., 2012; Harrison & Tong, 2009; Riggall & Postle, 2012; Serences et al., 2009; although this remains debated, see reviews Constantinidis et al., 2018; Lundqvist, Herman, & Miller, 2018) and in other task-relevant regions (e.g., visual and parietal cortex in visual WM tasks: Christophel et al., 2012; Harrison & Tong, 2009; Jerde et al., 2012; Rahmati et al., 2018; Riggall & Postle, 2012; Serences et al., 2009, reviewed in Christophel et al., 2017). Though the exact characterization of WM in the brain is not agreed on, most researchers agree that it is fundamentally different from other longer term memory processes, in that it requires active maintenance and is thus more fleeting, subject to decay, and more energy consuming.

Computational models of WM usually focus on either behavior or brain. For example, some models of WM behavior attempt to quantitatively characterize the nature of WM's limitations, and what this can teach us about the format of WM representations (e.g., Bays & Husain, 2008; Fougnie et al., 2012; Luck & Vogel, 1997; Nassar et al., 2018; van den Berg et al., 2012, 2014; Zhang & Luck, 2008). Models of the brain's WM signals attempt to characterize how stable but flexible representations can occur in biologically neural networks; they show that highly interconnected neural networks (e.g., some forms of *recurrent neural networks; RNNs*) can lead to stable attractor states that resemble the brain's neural activity during WM maintenance and account for behavior (Compte et al., 2000; Masse et al., 2019; Moody et al., 1998; Zipser, 1991). Despite these computational efforts and in contrast to RL, there is a less direct match of WM to a subfield of artificial intelligence. Some AI algorithms do include memory mechanisms to solve problems that cannot be solved by classic RL, because they require past information to be

maintained to make appropriate choices (called *Partially Observable Markov Decision Processes; POMDPs)*. These algorithms can share properties with biological WM, such as storing information in persistent activity rather than in network weights (e.g., RNNs), or maintain information over short periods of times in a controlled way where the agent can learn to gate the flow of information (e.g*., Long Short-Term Memory, LSTMs* (Hochreiter & Schmidhuber, 1997); Figure 1E). We will discuss the limitations of this similarity in the "Computational insights" section.

The interplay of RL and WM

The behavioral, neural, and computational instantiations within each process are overlapping, but distinct (see Eckstein et al., 2021 for a more in depth discussion on the distinctions in RL). For example, one can use "RL algorithms" to describe "RL behavior" (in cognitive modeling), and "RL algorithms" to explain "RL brain" (e.g., temporal difference learning well describes the striatal dopaminergic system in the brain). Similarly, "WM brain" is used for "WM behavior" (i.e., persistent activity in cortex might represent WM information that will be used to guide behavior).

However, these subfields can also be disjointed within one process, or can interact with the subfields of another process. For example, RL brain and WM brain can both recruit the same cortico-basal ganglial loop in the brain, suggesting there is less of a difference between "RL brain" and "WM brain" (expanded in "'RL and WM' brain"). Additionally, RL can help explain WM brain characteristics and how WM can selectively prioritize more behaviorally relevant information (expanded in "RL → WM"), and cognitive and neural WM processes can help describe RL behavior (expanded in "WM → RL"). These interactions across processes are important for understanding each process alone as well as behavior and the brain as a whole.

"RL and WM" brain

Reinforcement learning and working memory are often studied in isolation, and are often assumed to rely on predominantly different brain areas, at a first approximation. However, a closer examination shows that these processes are neurally and behaviorally tightly intertwined (Figure 1C). Frontal cortex (the "WM brain" area) and basal ganglia (the "RL brain" area) are

connected to one another through multiple parallel loops (Alexander et al., 1986; Haber, 2011). In addition to the frontal cortex and thalamus directly projecting onto one another, many parts of the cerebral cortex project onto the striatum, which then projects to the globus pallidus or substantia nigra pars reticulata, then to the thalamus, and back to the frontal cortex. These cortico-basal ganglia networks, traditionally studied in the motor control literature, have been demonstrated to be involved in both RL and WM tasks.

Prefrontal cortex has been implicated in many goal-directed RL tasks (Daw et al., 2005, 2011; Doll et al., 2016; Frank et al., 2007; Zhao et al., 2018), and activity in it has been shown to reflect reward prediction error (Javadi et al., 2014). Additionally, levels of dopamine in prefrontal cortex relate to WM performance (e.g., Bayram et al., 2021; Fallon et al., 2015; Fang et al., 2019). Damage to the basal ganglia can produce similar cognitive impairments as damage to the frontal cortex (e.g., Brown et al., 1997; Brown & Marsden, 1990; Middleton & Strick, 2000)**.** People with greater WM capacity / more WM resources are associated with better performance on serial response time tasks (de Kleijn et al., 2018), lower stress-induced detriments in instrumental behavioral tasks (Quaedflieg et al., 2019), learning acquisition (Segers et al., 2018), and lower biases in learning (Sidarta et al., 2018). Age-related RL decline may be due to not only decreased reward prediction error signalizing but also WM decline (van de Vijver & Ligneul, 2019).

The above only provides weak evidence in support of a possible overlap in WM and RL processes. Indeed, representations of information are distributed across the brain, and we often discuss brain specification for convenience, not because we believe one area to be necessary and sufficient for a type of task. Thus, the existence of overlapping neural and behavioral correlates of RL and WM is unsurprising and not a strong indicator of their interplay. In the remainder of this section, we provide more compelling and direct evidence of their interplay. Some successful, biologically realistic models of cortico-basal ganglial loops are able to account for WM prioritization and *WM gating*, where information is selectively allowed or "gated" into WM (Chatham et al., 2014; Chatham & Badre, 2015; Hazy et al., 2007; O'Reilly & Frank, 2006; expanded in "RL→ WM"). Depending on context, WM can contribute or interfere with RL processes in learning tasks (expanded in "WM → RL").

RL → WM

WM is the active maintenance of information in mind for later use. For example, when we are deciding what to wear in the morning, we may check our weather app, look out of the window to see the current weather, and remember the past days' weather (illustrated in Figure 2). When going to our closet to change, we must maintain these different sources of information in WM, weighting them based on their reliability for predicting today's weather. On one day, we may trust the current weather outside the most, dress inappropriately, and learn to rely on the weather app more in the future. In the future, we have learned to selectively maintain information from the weather app when getting dressed. How is information maintained in the brain over a working memory delay? How does the brain learn what information is important to remember? Here, we show that RL processes can contribute some answers to those questions. We will broadly discuss two families of models, the first concerning how information is maintained by the brain over a WM delay (which we will call *storage models*), and the second about how information is selectively maintained for goal-relevant behavior (which we will call *action models*).

Storage models

In storage tasks, participants remember some number of stimuli over a delay, then make a simple decision based on it. For example, participants may view a cluster of moving dots, maintain the dots' direction of motion over a delay, see another cluster of dots, and make a decision whether the motion is the same as before the delay.

RL algorithms for WM brain. RL models of WM storage tasks are mainly concerned with how neural signatures of WM may be reproduced. (Zipser, 1991) demonstrated that a recurrent learning-based neural network was able to capture the canonical elevated neural activity during the WM delay period, although using a biologically implausible backpropagation-through-time algorithm.

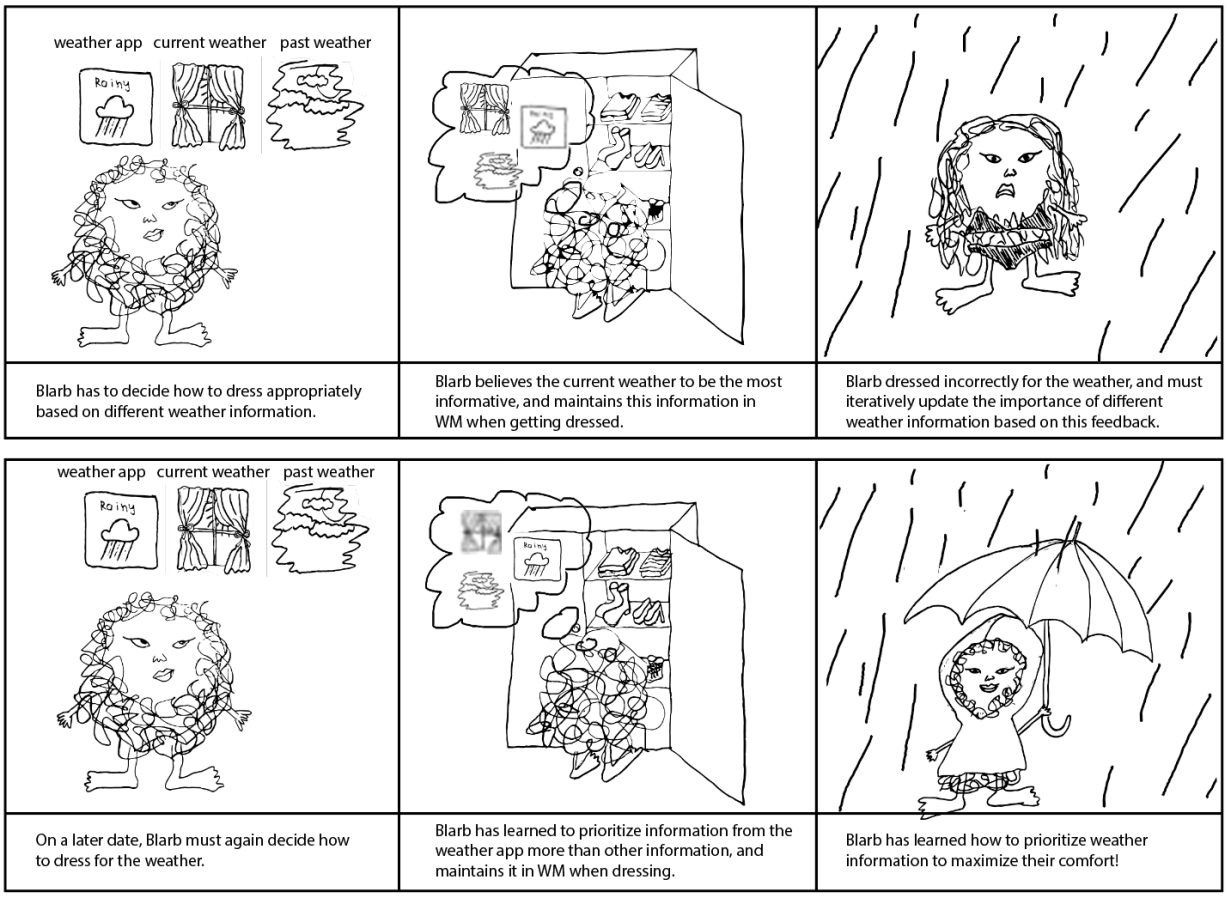

Figure 2. An illustrative example of how the RL process can be useful when learning what information to maintain in WM. When looking through our closet, we must maintain information relevant to today's weather in WM. RL provides an explanation of how we learn iteratively over time what information is relevant to maintain in WM for later decisions.

RL brain for WM behavior. If RL is involved in learning to use WM, can we see evidence of a role of dopamine in WM? Although dopamine does not exclusively represent RL in the brain (Lerner et al., 2021), it is strongly associated with RL processes, and is thus a reasonable heuristic for RL in the brain. The effects of dopamine levels on performance in WM tasks, while complex, are well characterized. For example, the role of dopamine in WM depends on task context (Furman et al., 2020), such as whether the task is spatial in nature (Gruszka et al., 2016; Luciana & Collins, 1997). It is debated whether it affects specifically interference / WM gating (Chatham et al., 2014; Chatham & Badre, 2015; Fallon, Mattiesing, et al., 2017; Fallon & Cools,

2014; Hazy et al., 2007; O'Reilly & Frank, 2006) or the precision with which one remembers items (Fallon, Zokaei, et al., 2017; Luciana et al., 1992). These differences could potentially be teased apart based on which dopaminergic system is being affected (striatal vs. frontal, antagonist vs agonist), the participant population (or more specifically, the baseline dopamine levels), and task design (some effects reviewed in (Cools & D'Esposito, 2011).

Action models

While storage models solve the simplest types of WM problems, where one or two stimuli need to be maintained across a WM delay, action models consider more realistic behavioral contexts, in which people are constantly presented with irrelevant information and WM must selectively process and store information. How does one learn what information is important to store in WM? In our real life example (Figure 2), how does one learn to selectively maintain information from the weather app when getting dressed?

This working memory gating process, the ability to selectively maintain a subset of incoming information, can be studied experimentally through dynamic choice tasks, such as the *1-2-AX task* (Frank et al., 2001; Figure 1D). In a simpler version of this task, the AX task, participants view a series of numbers and letters presented in time sequentially, and are instructed to respond with one key anytime A is directly followed by X, or when B is directly followed by Y. In the sequence 1, Y, A, X, 2, A, X, B, Y, the participant should respond on the fourth, seventh, and last trial. In the 1-2-AX task, there is an additional level of complexity, such that AX sequences should only be responded to when the most recent number was a 1, and BY sequences should only be responded to when the most recent number was a 2. In the above sequence, the participant should respond on only the fourth and last trial.

This task has several, nontrivial WM demands in order to optimize behavior. The participant must simultaneously evaluate incoming information (on a trial-to-trial level), selectively maintain information (e.g., the most recent context 1, then As and Xs), and rapidly update goals (e.g., a 2 is presented, then Bs and Ys must be maintained selectively). People are able to do this successfully, but how does WM learn what information to remember and what not to remember?

"Who" decides what information is selectively gated into WM? RL provides an explanation of how WM learns what information is important, and thus when to gate.

RL algorithm and brain for WM brain and behavior. Two examples of learning models of WM that include the ability for a WM process to maintain and update multiple items independently are the Long-Short-Term Memory (LSTM; Hochreiter & Schmidhuber, 1997) and *Prefrontal cortex - Basal ganglia Working Memor*y (*PBWM;* Hazy et al., 2007; O'Reilly & Frank, 2006) models. LSTM models were a computational innovation for neural network models. In addition to storing information in learned weights between neuron units, they also keep past information integrated into the network activity by feeding back past activity as an input to the current activity of neuron units. Their architecture utilized "memory blocks" with input, output, and forget gates (illustrated in Figure 1C) that allowed the network to independently and selectively maintain a number of stimuli, and maintain this information for much longer time periods than standard recurrent neural networks were able to do. The PBWM model was inspired by the connections between prefrontal cortex and basal ganglia, and offered a more biologically realistic model of how goal-relevant WM maintenance is learned. In this model, the basal ganglia learns through RL what is task relevant, and sends a teaching signal to prefrontal cortex which gates information in and out of memory. This model provided a critical extension from previous models (which include teaching signals from the basal ganglia to the prefrontal cortex; Braver & Cohen, 2000; Hochreiter & Schmidhuber, 1997) by articulating *how* the basal ganglia "knows" what is task relevant, and has been empirically supported (Rac-Lubashevsky & Frank, 2021). Both of these models are able to solve POMDPs, such as 1-2-AX by selectively storing the required information in memory.

These models have served as an inspiration for later models that either trade complexity for interpretability or adjust internal representations for improved task generalizability. For example Todd, Niv, Cohen (2009) replaced the biologically realistic neural network combination of RL, supervised learning, and unsupervised learning methods in PBWM with a simpler, tabular version of an RL algorithm, demonstrating the core functionality of the gating component within PBWM, without the complexity (but losing most of the biological realism). Another notable model, Working Memory Through Attentional Tagging (WorkMATe; Kruijne et al., 2021),

combines the simple, biologically-plausible learning algorithm as the AuGMEnT model (Rombouts et al., 2012; Rombouts, Bohte, & Roelfsema, 2015), the gating structures in LSTM and PBWM models, and abstract stimulus representations. While WorkMATe takes longer to train initially on tasks compared to other models, it is ultimately able to complete a broader range of tasks (including the 1-2-AX task) with more flexibility; furthermore, it generalizes better across previously unobserved stimuli and task modifications compared to Todd et al.'s altered PBWM model.

While these studies capture certain aspects of RL and WM, in a biologically-realistic way, they fail to capture all aspects. First, they fail to capture RL in a realistic way; these efforts typically use RL over very long time scales to train a network to do WM tasks, and in that sense do not relate to human RL (which is on a shorter time scale). Similarly, these models do not incorporate the limited-capacity of WM. How would these biologically-realistic models behave in scenarios when the amount of information exceeds the storage capacity of the WM process? (Todd et al.'s model does test this, and finds that the model fails in a human-like way in an artificial grammar task.) Studying experimental scenarios in which information exceeds WM capacity allows us to truly study how WM can dynamically change according to behavioral demands. In humans, individual items maintained in WM are not maintained in an all-or-none fashion, but with variable precision (Fougnie et al., 2012; van den Berg et al., 2012), and this precision tracks with behavioral relevance (Bays, 2014; Emrich et al., 2017; Klyszejko et al., 2014; Yoo et al., 2018). With these imperfectly-remembered representations, additional questions arise like whether or not agents know how imprecise their memory is (in humans, they seem to, since confidence scales with accuracy (Fougnie et al., 2012; Honig et al., 2020; Li et al., 2021; Rademaker et al., 2012; Samaha & Postle, 2017; Suchow et al., 2017; Vandenbroucke et al., 2014; Yoo et al., 2018), and, if so, if they can use that knowledge for performance-relevant behavior (humans and monkeys seem to be able to: Devkar et al., 2017; Honig et al., 2020; Yoo et al., 2018, 2021). Neural networks that can represent uncertainty (Swan & Wyble, 2014) or probability distributions over representations (Soltani & Wang, 2010) seem like promising routes to investigate these questions, and can help us further understand how behavioral relevance, interference, and decay all contribute to our WM representations.

WM → RL

As discussed earlier, RL can refer to a set of behaviors, a family of computational problems and algorithms, or a network in the brain. In this section, we discuss how WM processes are essential to explaining RL behavior. We focus on two situations: those where WM supports RL and those where WM offers a "redundant" learning mechanism.

WM provides inputs to RL processes. One example of the necessity of WM in some behavioral RL paradigms are scenarios in which some source of short-term memory is required to represent all information needed to make good decisions (POMDPs). For example, in an experiment where the correct response depends on the current and past trial's stimuli (such as 1-2-AX or the task in da Silva et al., 2017), participants must maintain the previous trials' stimuli in WM to learn the task correctly.

In various situations, WM may play a supporting role to RL's computations, such as providing input to RL computations by providing other information. In POMDPs, WM maintains stimulus information to correctly infer the current state. WM, however, may hold more than just stimulus information. For example, WM may maintain reward information itself (deviating from traditional theories in which reward information is only stored by RL processes): in the PFC-BG model developed by (Zhao et al., 2018), dopaminergic signals update both basal ganglia and prefrontal cortex, where reward information is encoded and updated in WM. Similarly, recent imaging work shows that WM helps transform novel goal stimuli into a signal the brain interprets as rewards for learning (McDougle et al., 2021). WM may assist RL by representing more abstract task-relevant information, to allow for generalization across tasks (Williams & Phillips, 2020), or by effectively lowering the set of states or actions RL operates over by filtering out irrelevant state spaces (Rmus et al., 2021). It would be an interesting direction of future study to investigate whether WM filters state spaces through attentional processes (e.g., Radulescu et al., 2019; Niv, 2019) and/or indirectly through its storage constraints.

WM as a parallel learning process. While WM supports RL, in particular in environments where a memory of past information is necessary, it may also be useful in cases where the state is

sufficient to determine the correct choice (MDPs). For example, learning how to make dumplings (raviolis, samosas, and/or other dishes that involve surrounding ingredients with a relatively small amount of dough) involves learning the proper amount of stuffing to use (illustrated in Figure 3). If using purely an RL process to learn the optimal amount of stuffing, you may try some amount, realize you put too much, experience a negative reward prediction error, and iteratively use less stuffing until you find the optimal amount; you will eventually learn the correct amount of stuffing, but it may be a slow and iterative experience. With WM, you could simply remember how much stuffing you used in the last dumpling, remove the appropriate amount, and immediately learn the correct amount. Thus, while WM is not necessary to complete this task, it allows you to learn much more efficiently and quickly.

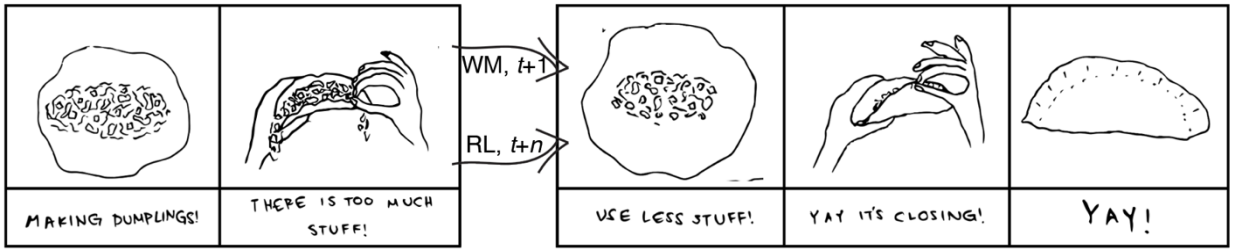

Figure 3. An illustrative example of how WM can be useful when learning. When learning how to make dumplings, one must learn the optimal amount of stuffing to put into the wrapper. If they put too much stuffing in for dumpling *t*, they can iteratively learn the correct amount of stuffing with an RL process (learning the correct amount in $n>>1$ trials), or immediately learn the correct amount using WM (on trial $t+1$). While using WM to maintain and manipulate information to calculate the correct amount of stuffing may be more effortful at first (from panel 2 to 3), it becomes effortless as the process is repeated.

This type of contribution of WM to RL behavior has been directly demonstrated in the stimulus-association task developed by Collins & Frank (2012). Within one block participants learned, through trial-and-error, the correct response to two to six different visual stimuli. Results showed that there was a classic WM set size effect on learning performance: participants were slower to learn on blocks where they had to learn the correct response of more stimuli in parallel. With two

stimuli, they appeared to make perfect use of memory by showing near optimal learning; with more stimuli, learning became increasingly more incremental. While purely RL models could not account for the results, even when improved with decay or interference mechanisms, they were well captured by a computational model where an RL and a capacity-limited WM component both contributed to choices, trading off depending on WM load. These results have been well replicated in the literature (Collins, 2018; Collins, Albrecht, et al., 2017; Collins et al., 2014; Jafarpour et al., 2019; McDougle & Collins, 2020; Viejo et al., 2015). In the "RLWM" model, the classic RL component learns iteratively the correct response associated with each stimulus from reward prediction errors. The WM component is implemented through an immediate-learning but decaying and capacity-limited process. This model has been modified by others. Viejo and others (2015) modeled working memory with a Bayesian working memory framework, such that previous trials' states, actions, and rewards are sampled to lower entropy until to some threshold (Viejo et al., 2015). This model as well as an extension of the RLWM model are able to capture reaction times across a range of phenomena (McDougle & Collins, 2020). In these models, WM and RL are essentially redundant; they both learn to represent state-action pairs (and succeed to varying degrees in different contexts); they are only identifiable in that they follow different dynamics (WM learns fast and forgets fast; RL learns slower but retains better).

Such RL+WM models have mostly treated them as independent processes that trade-off for choice. However, there is increasing evidence that this is an oversimplification, as the two processes appear to feed each other information. While RL and WM appear to cooperate during learning, this can lead to surprising competitive interference in the long-term retention of stimulus-response associations (Collins, Albrecht, et al., 2017; Collins, Ciullo, et al., 2017). Responses that were learned on blocks with lower set sizes, where WM would be sufficiently able to maintain all the necessary information, resulted in a higher detriment in performance during a later test phase compared to responses that were learned on blocks with higher set sizes, blocks in which WM alone would not have maintain all learned information (Collins, 2018). This "tortoise and hare" effect could also be seen in experimental paradigms manipulating study intervals (temporally-massed items vs spaced led to better relative performance during learning phase, but worse during later testing; Wimmer & Poldrack, 2020) and training context (blocked

context vs. interleaved led to better relative performance during learning phase, but worse during later testing; Shea & Morgan, 1979). This finding could be explained by an interaction of WM on the brain's RL mechanism, whereby WM fed reward expectations to the RL system, thus weakening the reward prediction error, and subsequent learning; an EEG study supported this theory by showing weaker RL-encoding neural signals in lower set sizes (Collins & Frank, 2018).

The importance of investigating inter-process dynamics

Considering how RL processes could affect participant performance in WM tasks is important when designing WM experiments and interpreting their results. For example, some WM studies investigate whether people *naturally* behave in a way that is consistent with an optimal Bayesian observer, showing that they already know how to use information (e.g., memory uncertainty; Keshvari et al., 2012; Yoo et al., 2021) to maximize performance and consequently don't have to learn it within the task. Because the RL literature has established that people can learn to behave optimally in relatively complex arbitrary tasks just from reinforcing feedback, it is important to either 1) withhold trial-to-trial feedback from the participants in these WM studies or 2) check for learning effects and interpret results accordingly (as reward is often used to motivate performance). Papers that implicitly argue that people behave optimally *naturally* but do not withhold correctness feedback and (e.g., Devkar et al., 2017; Honig et al., 2020; Yoo et al., 2018) could be misleading; people may be learning optimal behavior over the course of the experiment with an RL process.

Considering how WM contributes to behavior in RL tasks is equally important for making justified theoretical conclusions. For example, people with schizophrenia demonstrate deficits across a wide range of learning (Kim et al., 2007; Paulus et al., 2003) and RL tasks (Deserno et al., 2013; Gold et al., 2008), such as the Iowa Gambling Task (Shurman et al., 2005), probabilistic reinforcement, and reversal learning (Schlagenhauf et al., 2014; Waltz et al., 2007, 2011; Waltz & Gold, 2007) but not in all learning tasks (Deserno et al., 2013). Deficits in WM tasks (Barch & Ceaser, 2012), such as the Wisconsin Card Sorting test (Prentice et al., 2008) and change detection (Gold et al., 2003), are observed even more consistently. Studying either process in isolation may imply that schizophrenia affects both RL and WM processes. However,

Collins and others (2014) demonstrated that behavioral deficits in RL tasks in medicated people with schizophrenia could be entirely accounted for by WM's contribution to RL tasks. Indeed, once factoring out WM contributions, they observed no learning deficit (Collins, Albrecht, et al., 2017). This result could explain why some, but not other RL tasks lead to impairments, as they might have recruited WM differently. RLWM models can also account for age-related differences in behavior; the tortoise and hare effect changes with age, due to WM decline (van de Vijver et al., 2015; van de Vijver & Ligneul, 2019). These examples illustrate the risk of misattributing individual differences to the RL process when not accounting for potential other processes, such as WM.

There may be some hesitation to accept that there are two dissociable processes that are redundant (albeit computationally distinct). However, this redundancy is not unusual in other systems (e.g., multiple retinotopic maps spanning low-level visual to prefrontal brain areas) or even within RL. For example, there are separate dopaminergic systems in prefrontal cortex and striatum, and three different dopamine genes (two indexing striatal function and one prefrontal cortex function) have been behaviorally dissociated, such that slower reinforcement and avoidance behavior are related to the striatal genes, and a quicker recency related behavior associated with prefrontal gene (Frank et al., 2007). More recently, the RL field has widely focused on differently defined RL computations: a model-free RL which simply integrates value from past reward prediction errors, and a model-based RL which uses more knowledge about the environment to make more forward-looking decisions (Daw et al., 2005, 2011; Daw & O'Doherty, 2014; Dolan & Dayan, 2013). This dissociation has also been mapped onto individual differences in dopaminergic genetic polymorphisms, where model-free RL related more to striatal and model-based more to prefrontal function (Doll et al., 2016). While these dichotomies all have limitations (Collins & Cockburn, 2020), they illustrate the prevalence of partially-redundant systems. Exactly identifying these processes (e.g., how WM relates to model-based RL) is an important question for future research.

Considering the trade off between RL and WM processes in different environments may help us understand other behavior. For example, RL learning rates, as inferred from participants' behavior, increase in environments with more volatile reward structures (Behrens et al., 2007;

Iglesias et al., 2013). This behavior has been justified under a Bayesian framework, such that learning rates should increase with increasing uncertainty, which should increase with increasing environmental volatility (Courville et al., 2006; Mathys et al., 2011; Piray & Daw, 2020). A computational model that only considers a single RL process may find that the learning rate changes across contexts, but an RL+WM model may provide an alternative explanation for these results. Volatile environments may not result in an increase in the learning rate of the RL process, but lower the contribution of the temporally-slow RL process compared to the quick learning WM process. This interpretation is consistent with RL theories that suggest that a "model-based" process would be used more than a "model-free" process in higher-uncertainty situations (e.g., Daw et al., 2005; Pezzulo et al., 2013).

Interactions with other processes

While the purpose of this review is to specifically discuss the relationship between RL and WM processes, and the importance of studying them together, we would be remiss if we did not spend any time also discussing how RL and WM are affected by other processes like attention, episodic memory, and semantic memory. (While we do not discuss it here, we acknowledge that other processes are themselves influenced by WM and RL. attention: Downing, 2000; Olivers et al., 2006; Soto et al., 2005; Wilson & Niv, 2012; long-term memory: Ranganath et al., 2005; Shohamy & Adcock, 2010; motor action choice: Codol et al., 2018; Holland et al., 2018).

Attention has an immense effect on WM and RL, allowing us to filter information before storing it in WM (Chun et al., 2011; Souza et al., 2018) or learning from it (Farashahi et al., 2017; Leong et al., 2017; Niv et al., 2015). Brain areas associated with attentional control are similar to that of WM and RL (Braver et al., 2003; Dove et al., 2000; Leber et al., 2008). Some computational models of RL and WM explicitly include attention into the model. For example, in a modification of the ACT-R model (Anderson, 2007), attentional allocation is learned through RL, which informs what information should be held in WM (Stocco, 2017). This model is inspired by the cortico-basal ganglial loops, finding a relationship between behavioral measures of the indirect pathway in the basal ganglia and attention. Womelsdorf and others (2020) created a model with RL and WM components in addition to a selective suppression of non-chosen feature values and meta-learning mechanism adjusting exploration rates based on memory trace

of recent errors. These add ons are important to capture data in high-attentional load experimental conditions.

In addition to attention, other longer-term forms of memory like episodic and semantic memory affect WM and RL tasks. For example, episodic memories can disrupt working memory representations (Hoskin et al., 2019). Recent trial information or goals (Destefano et al., 2020; Kong et al., 2020) and global prior information (Destefano et al., 2020; Honig et al., 2020) also affect behavior on WM tasks. Episodic memory of previous choices on stimulus affects current choice in a learning task (Bornstein & Norman, 2017). Counterfactual learning of items chosen against one another is modulated by the strength of the episodic memory for them (Biderman & Shohamy, 2021). In some people, memory strength and RL learning rate seem to trade off depending on experimental learning context (Yifrah et al., 2021).

This relationship between long term memory and RL isn't particularly surprising, considering the importance a long-term storage would have on more realistic environments, which have high-dimensional, continuous, and partially-observable state spaces. In these scenarios, some other form of knowledge is required to approximate value functions over states that haven't been observed before, and over time-lengths between action and rewards that aren't realistic with an RL process alone. There has been an increasing effort to incorporate methods like "episodic learning" (RL augmented with episodic memory; reviewed in Gershman & Daw, 2017) and "experience replay" (computationally inspired from hippocampal replay, e.g., Foster & Wilson, 2006, using long-term memories of experiences to augment learning; e.g., Mnih et al. 2015; Lin, 1993) to achieve learning in more complex, realistic scenarios (e.g., Liu et al, 2021).

Computational models, such as the one developed by (Balkenius et al., 2020), investigate how attention, semantic memory, and episodic memory jointly affect decisions in addition to RL and WM. An example they provide is deciding between two pasta brands at the grocery store; there are a number of current features you can use to decide (price, packaging, ingredients), but information not currently observable (e.g., your memory of using a similar product, your knowledge that one is associated with a fancy restaurant) also affect your decision when choosing. This model seems to be fairly flexible, and can account for a variety of choices and

reaction times; empirical studies are necessary to investigate the ability of this model to account for real data. The importance of many interacting processes is represented in other models: attentional allocation informs which long-term memory representations should enter WM (Stocco, 2017); attention provides a solution to long-term credit assignment problems (Kruijne, 2021); and perception, working memory, and long-term memory contribute to rational decision making (Momennejad, 2021). Additionally, models like Todd et al.'s are considered WM-like, but are arguably closer to a form of long-term memory (Todd et al., 2009). Clearly, both long-term and working memory are important in ecological decision-making tasks, and stating the presence of such processes is important.

Just like how considering RL and WM in a vacuum neglects their complex interplay, considering these two processes alone also ignores their relationships with other processes. We believe investigating interactions between different complex processes is a difficult, but necessary challenge to understand the complexity of human behavior.

Computational insights

Reinforcement learning has long bridged Cognition and Computation, representing important parts of both modern AI research and psychology of learning and decision making, and showing how the two fields can be profitable sources of inspiration to each other. By contrast, working memory is acknowledged as an important aspect of human intelligence (Bull et al., 2008; Conway et al., 2003; Daneman & Carpenter, 1980; Harrison et al., 2015; Süß et al., 2002), but is a much less studied part of modern AI. Here, we explore AI's efforts to incorporate WM-like processes into learning agents, and discuss whether any computational insights could be gained by more cross-talk between cognition and AI in this domain.

Augmenting artificial agents with memory has long been recognized as necessary in some environments (Peshkin et al., 2001). In non-Markovian environments (e.g., POMDPs), the observable state is insufficient to determine an agent's policy, and keeping memory of past information allows the agent to create a new, more complex "latent state" that fully characterizes what choice should be taken. Originally, this form of memory has been set up as a lookup table of discrete events (Peshkin et al., 2001; Todd et al., 2009), being able to store an arbitrary

amount of information over arbitrary periods of time, and for this reason is often considered more related to episodic memory. Recent research in deep learning has successfully incorporated such additional memory processing to deep-RL agents (Botvinick et al., 2019; Duan et al., 2017; Graves et al., 2014; Nagabandi et al., 2019), further augmenting their capabilities and making them more flexible, for example enabling few-shot learning and imitation learning.

However, such memory mechanisms are not WM-like in multiple ways - they are potentially illimited, and rely on weight-like storage, rather than activity-based maintenance of information. There has been a recent push toward developing algorithms that have a human WM-like flexibility/generalizability across tasks, without focusing on biological realism. A promising avenue, called *meta-RL* or RL^2 (Botvinick et al., 2019; Duan et al., 2016; Wang et al., 2018), takes inspiration from meta-learning and uses slow RL algorithms to train deep neural networks with recurrent units to store information in such a way that the network's behavior (once its weights are fully trained and weights are frozen) mimics RL behavior at a fast, animal-like learning scale - i.e., uses recent reward information to make subsequent choices. This practice allows for neural networks to learn not only how to behave in one task (e.g., how to get the highest reward in a two-arm bandit task), but how to generalize its learning across similar tasks (e.g., all bandit tasks). These networks, however, also suffer from an initial training that is very slow and biologically implausible, which diminishes the viability of these models to explain human or animal learning. However, earlier efforts in simpler architectures have faster RL training times (Lloyd et al., 2012; Stocco, 2017), and show the potential fruitfulness of such approaches as models of human and animal learning in a more interpretable and biologically realistic way.

WM-like mechanisms in AI and theoretical neuroscience still diverge from what we know about how WM is used and implemented in humans in one very important way: biological WM is fundamentally resource limited. When optimizing just for precision of WM representations, having a limited WM capacity may seem like a bug. However, a capacity-limited WM is a feature when considering the metabolic costs for a biological agent (Musslick & Cohen, 2021; van den Berg & Ma, 2018). It can also be considered a feature because it forces humans' cognition to find the best compressed representations of their environment. For example, higher-

order statistics (Brady & Tenenbaum, 2013; Brady et al., 2009; Brady & Alvarez, 2011) and similarities across stimuli (Nassar et al., 2018) are used to compress WM representations such that we can (introduce some biases in representations but ultimately) remember "more." A capacity-limited WM allows for local, dynamic, efficient computation, with minimal practical effects on behavioral performance. Similarly, training artificial agents can come at a considerable environmental, computational, and financial cost (Hao, 2019; Strubell et al., 2019), and it may be beneficial to implement a limited-capacity WM process that can flexibly and dynamically allocate resources where behaviorally relevant.

In summary, while the AI field has usefully incorporated memory processes that share features with biological WM, none really captures the core of WM. Furthermore, those models that incorporate both RL and memory usually do not use RL at a time-scale that can be considered realistic in comparison to either RL brain or RL behavior. We hypothesize that AI might benefit from considering a human-like working mechanism, augmenting other learning and memory processes, to capture more human-like flexible learning and decision making in dynamic environments.

Conclusions

We aimed to review critical literature demonstrating the importance and interconnectedness of the RL and WM processes. The goal of this review is not to diminish the extremely important work done by those in both fields of RL and WM, but emphasize the importance of collaborating and considering how different processes affect one another. Keeping other processes in mind will allow us to make better experimental designs, make more general conclusions, and ultimately learn more about behavior and the brain. We believe it is of particular importance for the RL field to consider WM in their experiments, since even the simplest of learning tasks, usually thought to target only RL, have been shown to rely on WM processes (Collins & Frank, 2012; Frank et al., 2007; McDougle & Collins, 2020; Rmus et al., 2021). The continued study of RL and WM processes together will help us better understand the dynamics between them, the role of either in isolation, and behavior and the brain as a whole.

References

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195324259.001.0001

Baddeley, A. D., & Hitch, G. (1974). Working Memory. *The Psychology of Learning and Motivation*, *8*, 47–89. https://doi.org/10.1016/S0079-7421(08)60452-1

Balkenius, C., Tjøstheim, T. A., Johansson, B., Wallin, A., & Gärdenfors, P. (2020). The Missing Link Between Memory and Reinforcement Learning. *Frontiers in Psychology*, *11*, 560080. https://doi.org/10.3389/fpsyg.2020.560080

Barch, D. M., & Ceaser, A. (2012). Cognition in schizophrenia: Core psychological and neural mechanisms. *Trends in Cognitive Sciences*, *16*(1), 27–34. https://doi.org/10.1016/j.tics.2011.11.015

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*(1), 129–141. https://doi.org/10.1016/j.neuron.2005.05.020

Bayram, E., Litvan, I., Wright, B. A., Grembowski, C., Shen, Q., & Harrington, D. L. (2021). Dopamine effects on memory load and distraction during visuospatial working memory in cognitively normal Parkinson's disease. *Aging, Neuropsychology, and Cognition*, *28*(6), 812–828. https://doi.org/10.1080/13825585.2020.1828804

Bays, P. M. (2014). Noise in Neural Populations Accounts for Errors in Working Memory. *Journal of Neuroscience*, *34*(10), 3632–3645. https://doi.org/10.1523/JNEUROSCI.3204-13.2014

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851–854. https://doi.org/10.1126/science.1158023

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Bellemare, M. G., Candido, S., Castro, P. S., Gong, J., Machado, M. C., Moitra, S., Ponda, S. S., & Wang, Z. (2020). Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, *588*(7836), 77–82. https://doi.org/10.1038/s41586-020-2939-8

Biderman, N., & Shohamy, D. (2021). Memory and decision making interact to shape the value of unchosen options. *Nature Communications*, *12*, 4648. https://doi.org/10.1038/s41467-021-24907-x

Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience*, *20*(7), 997–1003. https://doi.org/10.1038/nn.4573

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement Learning, Fast and Slow. *Trends in Cognitive Sciences*, *23*(5), 408–422. https://doi.org/10.1016/j.tics.2019.02.006

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384–392. https://doi.org/10.1177/0956797610397956

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*(4), 487–502. https://doi.org/10.1037/a0016797

Brady, T. F., & Tenenbaum, B. J. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, *120*(1), 85–109. https://doi.org/10.1037/a0030779

Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working. *MIT Press. Making Working Memory Work*, 551–581.

Braver, T. S., Reynolds, J. R., & Donaldson, D. I. (2003). Neural mechanisms of transient and sustained cognitive control during task switching. *Neuron*, *39*(4), 713–726. https://doi.org/10.1016/s0896-6273(03)00466-5

Brown, J. (1958). Some Tests of the Decay Theory of Immediate Memory. *Quarterly Journal of Experimental Psychology*, *10*(1), 12–21. https://doi.org/10.1080/17470215808416249

Brown, L. L., Schneider, J. S., & Lidsky, T. I. (1997). Sensory and cognitive functions of the basal ganglia. *Current Opinion in Neurobiology*, *7*(2), 157–163. https://doi.org/10.1016/s0959-4388(97)80003-7

Brown, R. G., & Marsden, C. D. (1990). Cognitive function in Parkinson's disease: From description to theory. *Trends in Neurosciences*, *13*(1), 21–29. https://doi.org/10.1016/0166-2236(90)90058-i

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-Term Memory, Working Memory, and Executive Functioning in Preschoolers: Longitudinal Predictors of Mathematical Achievement at Age 7 Years. *Developmental Neuropsychology*, *33*(3), 205–228. https://doi.org/10.1080/87565640801982312

Chatham, C. H., & Badre, D. (2015). Multiple gates on working memory. *Current Opinion in Behavioral Sciences*, *1*, 23–31. https://doi.org/10.1016/j.cobeha.2014.08.001

Chatham, C. H., Frank, M. J., & Badre, D. (2014). Corticostriatal Output Gating during Selection from Working Memory. *Neuron*, *81*(4), 930–942. https://doi.org/10.1016/j.neuron.2014.01.002

Christophel, T. B., Hebart, M. N., & Haynes, J.-D. (2012). Decoding the Contents of Visual Short-Term Memory from Human Visual and Parietal Cortex. *Journal of Neuroscience*, *32*(38), 12983–12989. https://doi.org/10.1523/JNEUROSCI.0184-12.2012

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J.-D. (2017). The Distributed Nature of Working Memory. *Trends in Cognitive Sciences*, *21*(2), 111–124. https://doi.org/10.1016/j.tics.2016.12.007

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*, 73–101. https://doi.org/10.1146/annurev.psych.093008.100427

Codol, O., Holland, P. J., & Galea, J. M. (2018). The relationship between reinforcement and explicit control during visuomotor adaptation. *Scientific Reports*, *8*(1), 9121. https://doi.org/10.1038/s41598-018-27378-1

Collins, A. G. E. (2018). The Tortoise and the Hare: Interactions between Reinforcement Learning and Working Memory. *Journal of Cognitive Neuroscience*, *30*(10), 1422–1432. https://doi.org/10.1162/jocn_a_01238

Collins, A. G. E., Albrecht, M. A., Waltz, J. A., Gold, J. M., & Frank, M. J. (2017). Interactions Among Working Memory, Reinforcement Learning, and Effort in Value-Based Choice: A New Paradigm and Selective Deficits in Schizophrenia. *Biological Psychiatry*, *82*(6), 431–439. https://doi.org/10.1016/j.biopsych.2017.05.017

Collins, A. G. E., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working memory contributions to reinforcement learning impairments in schizophrenia. *Journal of Neuroscience*, *34*(41), 13747–13756. https://doi.org/10.1523/JNEUROSCI.0989-14.2014

Collins, A. G. E., Ciullo, B., Frank, M. J., & Badre, D. (2017). Working Memory Load Strengthens Reward Prediction Errors. *Journal of Neuroscience*, *37*(16), 4332–4342. https://doi.org/10.1523/JNEUROSCI.2700-16.2017

Collins, A. G. E., & Cockburn, J. (2020). Beyond simple dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, *21*(10), 576–586. https://doi.org/10.1038/s41583-020-0355-6

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis: Working memory in reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1024–1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x

Collins, A. G. E., & Frank, M. J. (2018). Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proceedings of the National Academy of Sciences*, *115*(10), 2502–2507. https://doi.org/10.1073/pnas.1720963115

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X. J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, *10*(9), 910–923. https://doi.org/10.1093/cercor/10.9.910

Constantinidis, C., Funahashi, S., Lee, D., Murray, J. D., Qi, X.-L., Wang, M., & Arnsten, A. F. T. (2018). Persistent Spiking Activity Underlies Working Memory. *Journal of Neuroscience*, *38*(32), 7020–7028. https://doi.org/10.1523/JNEUROSCI.2486-17.2018

Conway, Andrew R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547–552. https://doi.org/10.1016/j.tics.2003.10.005

Cools, R., & D'Esposito, M. (2011). Inverted-U–Shaped Dopamine Actions on Human Working Memory and Cognitive Control. *Biological Psychiatry*, *69*(12), e113–e125. https://doi.org/10.1016/j.biopsych.2011.03.028

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294–300. https://doi.org/10.1016/j.tics.2006.05.004

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*(4), 1158–1170. https://doi.org/10.3758/s13423-016-1191-6

da Silva, C. F., Yao, Y.-W., & Hare, T. A. (2017). Can model-free reinforcement learning operate over information stored in working-memory? *BioRxiv*. https://doi.org/10.1101/107698

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. https://doi.org/10.1038/nn1560

Daw, N. D., & O'Doherty, J. (2014). *Multiple Systems for Value Learning*. https://doi.org/10.1016/B978-0-12-416008-8.00021-8

Daw, N. D., & Tobler, P. N. (2014). Chapter 15 - Value Learning through Reinforcement: The Basics of Dopamine and Reinforcement Learning. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics (Second Edition)* (pp. 283–298). Academic Press. https://doi.org/10.1016/B978-0-12-416008-8.00015-2

de Kleijn, R., Kachergis, G., & Hommel, B. (2018). *IQ and working memory predict plan-based sequential action learning*. 6.

Deserno, L., Boehme, R., Heinz, A., & Schlagenhauf, F. (2013). Reinforcement Learning and Dopamine in Schizophrenia: Dimensions of Symptoms or Specific Features of a Disease Group? *Frontiers in Psychiatry*, *4*, 172. https://doi.org/10.3389/fpsyt.2013.00172

Destefano, I., Vul, E., & Brady, T. F. (2020). *Influences of both prior knowledge and recent history on visual working memory*. https://doi.org/10.31234/osf.io/ktrsj

Devkar, D., Wright, A. A., & Ma, W. J. (2017). Monkeys and humans take local uncertainty into account when localizing a change. *Journal of Vision*, *17*(11). https://doi.org/10.1167/17.11.4

Dolan, R. J., & Dayan, P. (2013). Goals and Habits in the Brain. *Neuron*, *80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

Doll, B. B., Bath, K. G., Daw, N. D., & Frank, M. J. (2016). Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning. *Journal of Neuroscience*, *36*(4), 1211–1222. https://doi.org/10.1523/JNEUROSCI.1901-15.2016

Dove, A., Pollmann, S., Schubert, T., Wiggins, C. J., & von Cramon, D. Y. (2000). Prefrontal cortex activation in task switching: An event-related fMRI study. *Brain Research. Cognitive Brain Research*, *9*(1), 103–109. https://doi.org/10.1016/s0926-6410(99)00029-4

Downing, P. E. (2000). Interactions between visual working memory and selective attention. *Psychological Science*, *11*(6), 467–473. https://doi.org/10.1111/1467-9280.00290

Duan, Y., Andrychowicz, M., Stadie, B. C., Ho, J., Schneider, J., Sutskever, I., Abbeel, P., & Zaremba, W. (2017). One-Shot Imitation Learning. *ArXiv*. http://arxiv.org/abs/1703.07326

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016). RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning. *ArXiv*. http://arxiv.org/abs/1611.02779

Eckstein, M. K., Wilbrecht, L., & Collins, A. G. E. (2021). What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. *Current Opinion in Behavioral Sciences*, *41*, 128–137. https://doi.org/10.1016/j.cobeha.2021.06.004

Emrich, M. S., Lockhart, A. H., & Al-Aidroos, N. (2017). Attention Mediates the Flexible Allocation of Visual Working Memory Resources. *Journal of Experimental Psychology. Human Perception and Performance*. https://doi.org/10.1037/xhp0000398

Fallon, S. J., & Cools, R. (2014). Reward acts on the pFC to enhance distractor resistance of working memory representations. *Journal of Cognitive Neuroscience*, *26*(12), 2812–2826. https://doi.org/10.1162/jocn_a_00676

Fallon, S. J., Mattiesing, R. M., Muhammed, K., Manohar, S., & Husain, M. (2017). Fractionating the Neurocognitive Mechanisms Underlying Working Memory: Independent Effects of Dopamine and Parkinson's Disease. *Cerebral Cortex (New York, N.Y.: 1991)*, *27*(12), 5727–5738. https://doi.org/10.1093/cercor/bhx242

Fallon, S. J., Smulders, K., Esselink, R. A., van de Warrenburg, B. P., Bloem, B. R., & Cools, R. (2015). Differential optimal dopamine levels for set-shifting and working memory in Parkinson's disease. *Neuropsychologia*, *77*, 42–51. https://doi.org/10.1016/j.neuropsychologia.2015.07.031

Fallon, S. J., Zokaei, N., Norbury, A., Manohar, S. G., & Husain, M. (2017). Dopamine Alters the Fidelity of Working Memory Representations according to Attentional Demands. *Journal of Cognitive Neuroscience*, *29*(4), 728–738. https://doi.org/10.1162/jocn_a_01073

Fang, Y. J., Tan, C. H., Tu, S. C., Liu, C.-Y., & Yu, R. L. (2019). More than an "inverted-U"? An exploratory study of the association between the catechol-o-methyltransferase gene polymorphism and executive functions in Parkinson's disease. *PLOS ONE*, *14*(3), e0214146. https://doi.org/10.1371/journal.pone.0214146

Farashahi, S., Rowe, K., Aslami, Z., Lee, D., & Soltani, A. (2017). Feature-based learning improves adaptability without compromising precision. *Nature Communications*, *8*(1), 1768. https://doi.org/10.1038/s41467-017-01874-w

Foster, D. & Wilson, M. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature,* 440**,** 680–683. https://doi.org/10.1038/nature04587

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature Communications*, *3*, 1229. https://doi.org/10.1038/ncomms2237

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, & Behavioral Neuroscience*, *1*(2), 137–160. https://doi.org/10.3758/CABN.1.2.137

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, *104*(41), 16311–16316. https://doi.org/10.1073/pnas.0706111104

Funahashi, S., Bruce, J. C., & Goldman-Rakic, S. P. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *61*(2), 331–349. https://doi.org/10.1152/jn.1989.61.2.331

Furman, D. J., White, R. L., III, Naskolnakorn, J., Ye, J., Kayser, A., & D'Esposito, M. (2020). Effects of Dopaminergic Drugs on Cognitive Control Processes Vary by Genotype. *Journal of Cognitive Neuroscience*, *32*(5), 804–821. https://doi.org/10.1162/jocn_a_01518

Fuster, M. J., & Alexander, E. G. (1971). Neuron activity related to short-term memory. *Science*, *173*(3997), 652–654. https://doi.org/10.1126/science.173.3997.652

Gershman, S. J., & Daw, N. D. (2017). Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annual Review of Psychology*, *68*(1), 101–128. https://doi.org/10.1146/annurev-psych-122414-033625

Gold, J. M., Waltz, J. A., Prentice, K. J., Morris, S. E., & Heerey, E. A. (2008). Reward processing in schizophrenia: A deficit in the representation of value. *Schizophrenia Bulletin*, *34*(5), 835–847. https://doi.org/10.1093/schbul/sbn068

Gold, J. M., Wilk, C. M., McMahon, R. P., Buchanan, R. W., & Luck, S. J. (2003). Working memory for visual features and conjunctions in schizophrenia. *Journal of Abnormal Psychology*, *112*(1), 61–71. https://doi.org/10.1037/0021-843X.112.1.61

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing Machines. *ArXiv*. http://arxiv.org/abs/1410.5401

Gruszka, A., Bor, D., Barker, R. R., Necka, E., & Owen, A. M. (2016). The role of executive processes in working memory deficits in Parkinson's disease. *Polish Psychological Bulletin*, *47*(1), 123–130. https://doi.org/10.1515/ppb-2016-0013

Haber, S. N. (2011). Neural Circuits of Reward and Decision Making: Integrative Networks across Corticobasal Ganglia Loops. In R. B. Mars, J. Sallet, M. F. S. Rushworth, & N. Yeung (Eds.), *Neural Basis of Motivational and Cognitive Control*. MIT Press.

Hao, K. (2019). *Training a single AI model can emit as much carbon as five cars in their lifetimes*. MIT Technology Review. https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632–635. https://doi.org/10.1038/nature07832

Harrison, T. L., Shipstead, Z., & Engle, R. W. (2015). Why is working memory capacity related to matrix reasoning tasks? *Memory & Cognition*, *43*(3), 389–396. https://doi.org/10.3758/s13421-014-0473-3

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: Computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1601–1613. https://doi.org/10.1098/rstb.2007.2055

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Holland, P., Codol, O., & Galea, J. M. (2018). Contribution of explicit processes to reinforcement-based motor learning. *Journal of Neurophysiology*, *119*(6), 2241–2255. https://doi.org/10.1152/jn.00901.2017

Honig, M., Ma, W. J., & Fougnie, D. (2020). Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(15), 8391–8397. https://doi.org/10.1073/pnas.1918143117

Hoskin, A. N., Bornstein, A. M., Norman, K. A., & Cohen, J. D. (2019). Refresh my memory: Episodic memory reinstatements intrude on working memory maintenance. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(2), 338–354. https://doi.org/10.3758/s13415-018-00674-z

Houk, J. C. (1995). Information Processing in Modular Circuits Linking Basal Ganglia and Cerebral Cortex. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia*. The MIT Press. https://doi.org/10.7551/mitpress/4708.003.0004

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E. M., & Stephan, K. E. (2013). Hierarchical Prediction Errors in Midbrain and Basal Forebrain during Sensory Learning. *Neuron*, *80*(2), 519–530. https://doi.org/10.1016/j.neuron.2013.09.009

Jafarpour, A., Buffalo, E. A., Knight, R. T., & Collins, A. G. E. (2019). Event segmentation reveals working memory forgetting rate. *BioRxiv*. https://doi.org/10.1101/571380

Javadi, A. H., Schmidt, D. H. K., & Smolka, M. N. (2014). Adolescents Adapt More Slowly than Adults to Varying Reward Contingencies. *Journal of Cognitive Neuroscience*, *26*(12), 2670–2681. https://doi.org/10.1162/jocn_a_00677

Jerde, A. T., Merriam, P. E., Riggall, C. A., Hedges, H. J., & Curtis, C. E. (2012). Prioritized maps of space in human frontoparietal cortex. *Journal of Neuroscience*, *32*(48), 17382–17390. https://doi.org/10.1523/JNEUROSCI.3810-12.2012

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, *15*(4–6), 535–547. https://doi.org/10.1016/S0893-6080(02)00047-3

Keshvari, S., van den Berg, R., & Ma, W. J. (2012). Probabilistic Computation in Human Perception under Variability in Encoding Precision. *PLoS ONE*, *7*(6), e40216. https://doi.org/10.1371/journal.pone.0040216

Kim, H., Lee, D., Shin, Y.-M., & Chey, J. (2007). Impaired strategic decision making in schizophrenia. *Brain Research*, *1180*, 90–100. https://doi.org/10.1016/j.brainres.2007.08.049

Klyszejko, Z., Rahmati, M., & Curtis, C. E. (2014). Attentional priority determines working memory precision. *Vision Research*, *105*, 70–76. https://doi.org/10.1016/j.visres.2014.09.002

Kong, G., Meehan, J., & Fougnie, D. (2020). Working memory is corrupted by strategic changes in search templates. *Journal of Vision*, *20*(8), 3. https://doi.org/10.1167/jov.20.8.3

Kruijne, W., Bohte, S. M., Roelfsema, P. R., & Olivers, C. N. L. (2021). Flexible Working Memory Through Selective Gating and Attentional Tagging. *Neural Computation*, *33*(1), 1–40. https://doi.org/10.1162/neco_a_01339

Leber, A. B., Turk-Browne, N. B., & Chun, M. M. (2008). Neural predictors of moment-to-moment fluctuations in cognitive flexibility. *Proceedings of the National Academy of Sciences*, *105*(36), 13592–13597. https://doi.org/10.1073/pnas.0805423105

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron*, *93*(2), 451–463. https://doi.org/10.1016/j.neuron.2016.12.040

Lerner, T. N., Holloway, A. L., & Seiler, J. L. (2021). Dopamine, Updated: Reward Prediction Error and Beyond. *Current Opinion in Neurobiology*, *67*, 123–130. https://doi.org/10.1016/j.conb.2020.10.012

Li, H.-H., Sprague, T. C., Yoo, A. H., Ma, W. J., & Curtis, C. E. (2021). Joint representation of working memory and uncertainty in human cortex. *Neuron*. https://doi.org/10.1016/j.neuron.2021.08.022

Lin, L.-J. Reinforcement learning for robots using neural networks. (1993). Technical Report, DTIC Document.

Liu Y, Mattar M.G., Behrens T.E.J., Daw N.D., & Dolan R.J. (2021). Experience replay is associated with efficient nonlocal learning. *Science.* 372(6544). https://doi.org/10.1126/science.abf1357

Lloyd, K., Becker, N., Jones, M. W., & Bogacz, R. (2012). Learning to use working memory: A reinforcement learning gating model of rule acquisition in rats. *Frontiers in Computational Neuroscience*, *6*. https://doi.org/10.3389/fncom.2012.00087

Luciana, M., & Collins, P. F. (1997). Dopaminergic Modulation of Working Memory for Spatial but Not Object Cues in Normal Humans. *Journal of Cognitive Neuroscience*, *9*(3), 330–347. https://doi.org/10.1162/jocn.1997.9.3.330

Luciana, M., Depue, R. A., Arbisi, P., & Leon, A. (1992). Facilitation of Working Memory in Humans by a D2 Dopamine Receptor Agonist. *Journal of Cognitive Neuroscience*, *4*(1), 58–68. https://doi.org/10.1162/jocn.1992.4.1.58

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281. https://doi.org/10.1038/36846

Lundqvist, M., Herman, P., & Miller, K. E. (2018). Working Memory: Delay Activity, Yes! Persistent Activity? Maybe Not. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *38*(32), 7013–7019. https://doi.org/10.1523/JNEUROSCI.2485-17.2018

Maes, E. J. P., Sharpe, M. J., Usypchuk, A. A., Lozzi, M., Chang, C. Y., Gardner, M. P. H., Schoenbaum, G., & Iordanova, M. D. (2020). Causal evidence supporting the proposal that dopamine transients function as temporal difference prediction errors. *Nature Neuroscience*, *23*(2), 176–178. https://doi.org/10.1038/s41593-019-0574-1

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X. J., & Freedman, D. J. (2019). Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature Neuroscience*, *22*(7), 1159–1167. https://doi.org/10.1038/s41593-019-0414-3

Mathys, C., Daunizeau, J., Friston, K., & Stephan, K. (2011). A Bayesian Foundation for Individual Learning Under Uncertainty. *Frontiers in Human Neuroscience*, *5*, 39. https://doi.org/10.3389/fnhum.2011.00039

McDougle, S. D., Ballard, I. C., Baribault, B., Bishop, S. J., & Collins, A. G. E. (2021). Executive Function Assigns Value to Novel Goal-Congruent Outcomes. *Cerebral Cortex*, *bhab205*. https://doi.org/10.1093/cercor/bhab205

McDougle, S. D., & Collins, A. G. E. (2020). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental

learning. *Psychonomic Bulletin & Review*, *28*(1), 20–39. https://doi.org/10.3758/s13423-020-01774-z

Middleton, F. A., & Strick, P. L. (2000). Basal Ganglia Output and Cognition: Evidence from Anatomical, Behavioral, and Clinical Studies. *Brain and Cognition*, *42*(2), 183–200. https://doi.org/10.1006/brcg.1999.1099

Mnih, V., Kavukcuoglu, K., Silver, D. *et al.* (2015). Human-level control through deep reinforcement learning. *Nature*, 518**,** 529–533. https://doi.org/10.1038/nature14236

Momennejad, I., Lewis-Peacock, J., Norman, K. A., Cohen, J. D., Singh, S., & Lewis, R. L. (2021). Rational use of episodic and working memory: A normative account of prospective memory. *Neuropsychologia*, *158*, 107657. https://doi.org/10.1016/j.neuropsychologia.2020.107657

Montague, P., Dayan, P., & Sejnowski, T. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience*, *16*(5), 1936–1947. https://doi.org/10.1523/JNEUROSCI.16-05-01936.1996

Moody, S. L., Wise, S. P., di Pellegrino, G., & Zipser, D. (1998). A Model That Accounts for Activity in Primate Frontal Cortex during a Delayed Matching-to-Sample Task. *The Journal of Neuroscience*, *18*(1), 399–410. https://doi.org/10.1523/JNEUROSCI.18-01-00399.1998

Murray, J. D., Bernacchia, A., Roy, N. A., Constantinidis, C., Romo, R., & Wang, X. J. (2017). Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proceedings of the National Academy of Sciences*, *114*(2), 394–399. https://doi.org/10.1073/pnas.1619449114

Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, *25*(9), 757–775. https://doi.org/10.1016/j.tics.2021.06.001

Nagabandi, A., Clavera, I., Liu, S., Fearing, R. S., Abbeel, P., Levine, S., & Finn, C. (2019). Learning to Adapt in Dynamic, Real-World Environments Through Meta-Reinforcement Learning. *ArXiv*. http://arxiv.org/abs/1803.11347

Nassar, R. M., Helmers, C. J., & Frank, J. M. (2018). Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological Review*, *125*(4), 486–511. https://doi.org/10.1037/rev0000101

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154. https://doi.org/10.1016/j.jmp.2008.12.005

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*(10), 1544–1553. https://doi.org/10.1038/s41593-019-0470-8

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885–958. https://doi.org/10.1037/bul0000153

Olivers, C. N. L., Meijer, F., & Theeuwes, J. (2006). Feature-based memory-driven attentional capture: Visual working memory content affects visual attention. *Journal of Experimental Psychology. Human Perception and Performance*, *32*(5), 1243–1265. https://doi.org/10.1037/0096-1523.32.5.1243

O'Reilly, R. C., & Frank, M. J. (2006). Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. *Neural Computation*, *18*(2), 283–328. https://doi.org/10.1162/089976606775093909

Paulus, M. P., Frank, L., Brown, G. G., & Braff, D. L. (2003). Schizophrenia Subjects Show Intact Success-Related Neural Activation but Impaired Uncertainty Processing during Decision-Making. *Neuropsychopharmacology*, *28*(4), 795–806. https://doi.org/10.1038/sj.npp.1300108

Peshkin, L., Meuleau, N., & Kaelbling, L. (2001). Learning Policies with External Memory. *ArXiv*. http://arxiv.org/abs/cs/0103003

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*(3), 193–198. http://dx.doi.org/10.1037/h0049234

Pezzulo, G., Rigoli, F., & Chersi, F. (2013). The Mixed Instrumental Controller: Using Value of Information to Combine Habitual Choice and Mental Simulation. *Frontiers in Psychology*, *4*, 92. https://doi.org/10.3389/fpsyg.2013.00092

Piray, P., & Daw, N. D. (2020). A simple model for learning in volatile environments. *PLOS Computational Biology*, *16*(7), e1007963. https://doi.org/10.1371/journal.pcbi.1007963

Prentice, K. J., Gold, J. M., & Buchanan, R. W. (2008). The Wisconsin Card Sorting impairment in schizophrenia is evident in the first four trials. *Schizophrenia Research*, *106*(1), 81–87. https://doi.org/10.1016/j.schres.2007.07.015

Quaedflieg, C. W. E. M., Stoffregen, H., Sebalo, I., & Smeets, T. (2019). Stress-induced impairment in goal-directed instrumental behaviour is moderated by baseline working memory. *Neurobiology of Learning and Memory*, *158*, 42–49. https://doi.org/10.1016/j.nlm.2019.01.010

Rac-Lubashevsky, R., & Frank, M. J. (2021). Analogous computations in working memory input, output and motor gating: Electrophysiological and computational modeling evidence. *PLOS Computational Biology*, *17*(6), e1008971. https://doi.org/10.1371/journal.pcbi.1008971

Rademaker, R. L., Tredway, C. H., & Tong, F. (2012). Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *Journal of Vision*, *12*(13), 21–21. https://doi.org/10.1167/12.13.21

Radulescu, A., Niv, Y., & Ballard, I. (2019). Holistic Reinforcement Learning: The Role of Structure and Attention. *Trends in Cognitive Sciences*, *23*(4), 278–292. https://doi.org/10.1016/j.tics.2019.01.010

Rahmati, M., Saber, G. T., & Curtis, C. E. (2018). Population Dynamics of Early Visual Cortex during Working Memory. *Journal of Cognitive Neuroscience*, *30*(2), 219–233. https://doi.org/10.1162/jocn_a_01196

Ranganath, C., Cohen, M. X., & Brozinsky, C. J. (2005). Working memory maintenance contributes to long-term memory formation: Neural and behavioral evidence. *Journal of Cognitive Neuroscience*, *17*(7), 994–1010. https://doi.org/10.1162/0898929054475118

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. In A.H. Black & W.F. Prokasy (Eds.). Classical Conditioning: Current Research and Theory. Appleton-Century-Crofts, NY.

Riggall, A. C., & Postle, B. R. (2012). The Relationship between Working Memory Storage and Elevated Activity as Measured with Functional Magnetic Resonance Imaging. *Journal of Neuroscience*, *32*(38), 12990–12998. https://doi.org/10.1523/JNEUROSCI.1892-12.2012

Rmus, M., McDougle, S. D., & Collins, A. G. E. (2021). The role of executive function in shaping reinforcement learning. *Current Opinion in Behavioral Sciences*, *38*, 66–73. https://doi.org/10.1016/j.cobeha.2020.10.003

Samaha, J., & Postle, B. R. (2017). Correlated individual differences suggest a common mechanism underlying metacognition in visual perception and visual short-term memory. *Proceedings of The Royal Society B Biological Sciences*, *284*(1867), 20172035. https://doi.org/10.1098/rspb.2017.2035

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of Action-Specific Reward Values in the Striatum. *Science*, *310*(5752), 1337–1340. https://doi.org/10.1126/science.1115270

Satoh, T., Nakai, S., Sato, T., & Kimura, M. (2003). Correlated Coding of Motivation and Outcome of Decision by Dopamine Neurons. *Journal of Neuroscience*, *23*(30), 9913–9923. https://doi.org/10.1523/JNEUROSCI.23-30-09913.2003

Schlagenhauf, F., Huys, Q. J. M., Deserno, L., Rapp, M. A., Beck, A., Heinze, H.-J., Dolan, R., & Heinz, A. (2014). Striatal dysfunction during reversal learning in unmedicated schizophrenia patients. *NeuroImage*, *89*, 171–180. https://doi.org/10.1016/j.neuroimage.2013.11.034

Schultz, W. (2002). Getting Formal with Dopamine and Reward. *Neuron*, *36*(2), 241–263. https://doi.org/10.1016/S0896-6273(02)00967-4

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*, 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schultz, W., Romo, R., Ljungberg, T., Mirenowicz, J., Hollerman, J. R., & Dickinson, A. (1995). Reward-related Signals Carried by Dopamine Neurons. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 233–248). MIT Press.

Segers, E., Beckers, T., Geurts, H., Claes, L., Danckaerts, M., & van der Oord, S. (2018). Working Memory and Reinforcement Schedule Jointly Determine Reinforcement Learning in Children: Potential Implications for Behavioral Parent Training. *Frontiers in Psychology*, *9*, 394. https://doi.org/10.3389/fpsyg.2018.00394

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-Specific Delay Activity in Human Primary Visual Cortex. *Psychological Science*, *20*(2), 207–214. https://doi.org/10.1111/j.1467-9280.2009.02276.x

Shea, J. B., & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(2), 179–187. https://doi.org/10.1037/0278-7393.5.2.179

Shohamy, D., & Adcock, R. A. (2010). Dopamine and adaptive memory. *Trends in Cognitive Sciences*, *14*(10), 464–472. https://doi.org/10.1016/j.tics.2010.08.002

Shurman, B., Horan, W. P., & Nuechterlein, K. H. (2005). Schizophrenia patients demonstrate a distinctive pattern of decision-making impairment on the Iowa Gambling Task. *Schizophrenia Research*, *72*(2–3), 215–224. https://doi.org/10.1016/j.schres.2004.03.020

Sidarta, A., van Vugt, F. T., & Ostry, D. J. (2018). Somatosensory working memory in human reinforcement-based motor learning. *Journal of Neurophysiology*, *120*(6), 3275–3286. https://doi.org/10.1152/jn.00442.2018

Soltani, A., & Wang, X. J. (2010). Synaptic computation underlying probabilistic inference. *Nature Neuroscience*, *13*(1), 112–119. https://doi.org/10.1038/nn.2450

Soto, D., Heinke, D., Humphreys, G. W., & Blanco, M. J. (2005). Early, involuntary top-down guidance of attention from working memory. *Journal of Experimental Psychology. Human Perception and Performance*, *31*(2), 248–261. https://doi.org/10.1037/0096-1523.31.2.248

Souza, A. S., Thalmann, M., & Oberauer, K. (2018). The precision of spatial selection into the focus of attention in working memory. *Psychonomic Bulletin & Review*, *25*(6), 2281–2288. https://doi.org/10.3758/s13423-018-1471-4

Sternberg, S. (1966). High-Speed Scanning in Human Memory. *Science*, *153*(3736), 652–654. https://doi.org/10.1126/science.153.3736.652

Stocco, A. (2017). An Integrated Computational Framework for Attention, Reinforcement Learning, and Working Memory. *AAAI 2017 Fall Symposium*, 6.

Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405. https://doi.org/10.1016/j.tics.2015.05.004

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *ArXiv*. http://arxiv.org/abs/1906.02243

Suchow, J. W., Fougnie, D., & Alvarez, G. A. (2017). Looking inward and back: Real-time monitoring of visual working memories. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *43*(4), 660–668. https://doi.org/10.1037/xlm0000320

Suri, R. E., Bargas, J., & Arbib, M. A. (2001). Modeling functions of striatal dopamine modulation in learning and planning. *Neuroscience*, *103*(1), 65–85. https://doi.org/10.1016/S0306-4522(00)00554-6

Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—And a little bit more. *Intelligence*, *30*(3), 261–288. https://doi.org/10.1016/S0160-2896(01)00100-3

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.

Swan, G., & Wyble, B. (2014). The binding pool: A model of shared neural resources for distinct items in visual working memory. *Attention, Perception & Psychophysics*, *76*(7), 2136–2157. https://doi.org/10.3758/s13414-014-0633-3

Todd, M. T., Niv, Y., & Cohen, J. D. (2009). Learning to Use Working Memory in Partially Observable Environments through Dopaminergic Reinforcement. *Advances in Neural Information Processing Systems*, *21*, 1689–1696. https://doi.org/10.1371/journal.pone.0075455

van de Vijver, I., & Ligneul, R. (2019). Relevance of working memory for reinforcement learning in older adults varies with timescale of learning. *Aging, Neuropsychology, and Cognition*, *27*(5), 654–676. https://doi.org/10.1080/13825585.2019.1664389

van de Vijver, I., Ridderinkhof, K. R., & de Wit, S. (2015). Age-related changes in deterministic learning from positive versus negative performance feedback. *Aging, Neuropsychology, and Cognition*, *22*(5), 595–619. https://doi.org/10.1080/13825585.2015.1020917

van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review*, *121*(1), 124–149. https://doi.org/10.1037/a0035234

van den Berg, R., & Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *ELife*, *7*. https://doi.org/10.7554/eLife.34963

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780–8785. https://doi.org/10.1073/pnas.1117465109

Vandenbroucke, E. A. R., Sligte, I. G., Barrett, A. B., Seth, A. K., Fahrenfort, J. J., & Lamme, V. A. F. (2014). Accurate metacognition for visual sensory memory representations. *Psychological Science*, *25*(4), 861–873. https://doi.org/10.1177/0956797613516146

Viejo, G., Khamassi, M., Brovelli, A., & Girard, B. (2015). Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00225

Waltz, J. A., Frank, M. J., Robinson, B. M., & Gold, J. M. (2007). Selective reinforcement learning deficits in schizophrenia support predictions from computational models of striatal-cortical dysfunction. *Biological Psychiatry*, *62*(7), 756–764. https://doi.org/10.1016/j.biopsych.2006.09.042

Waltz, J. A., Frank, M. J., Wiecki, T. V., & Gold, J. M. (2011). Altered probabilistic learning and response biases in schizophrenia: Behavioral evidence and neurocomputational modeling. *Neuropsychology*, *25*(1), 86–97. https://doi.org/10.1037/a0020882

Waltz, J. A., & Gold, J. M. (2007). Probabilistic reversal learning impairments in schizophrenia: Further evidence of orbitofrontal dysfunction. *Schizophrenia Research*, *93*(1–3), 296–303. https://doi.org/10.1016/j.schres.2007.03.010

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868. https://doi.org/10.1038/s41593-018-0147-8

Wickens, J. R. (2009). Synaptic plasticity in the basal ganglia. *Behavioural Brain Research*, *199*(1), 119–128. https://doi.org/10.1016/j.bbr.2008.10.030

Williams, A., & Phillips, J. (2020). Transfer Reinforcement Learning Using Output-Gated Working Memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(02), 1324–1331. https://doi.org/10.1609/aaai.v34i02.5488

Wilson, R., & Niv, Y. (2012). Inferring Relevance in a Changing World. *Frontiers in Human Neuroscience*, *5*, 189. https://doi.org/10.3389/fnhum.2011.00189

Wimmer, G. E., & Poldrack, R. A. (2020). Reward learning and working memory: Effects of massed versus spaced training and post-learning delay period. *BioRxiv*. https://doi.org/10.1101/2020.03.19.997098

Womelsdorf, T., Watson, M. R., & Tiesinga, P. (2020). Learning at variable attentional load requires cooperation between working memory, meta-learning and attention-augmented reinforcement learning. *BioRxiv*. https://doi.org/10.1101/2020.09.27.315432

Yifrah, B., Ramaty, A., Morris, G., & Mendelsohn, A. (2021). Individual differences in experienced and observational decision-making illuminate interactions between reinforcement learning and declarative memory. *Scientific Reports*, *11*(1), 5899. https://doi.org/10.1038/s41598-021-85322-2

Yoo, A. H., Acerbi, L., & Ma, W. J. (2021). Uncertainty is maintained and used in working memory. *Journal of Vision*, *21*(8), 13–13. https://doi.org/10.1167/jov.21.8.13

Yoo, A. H., Klyszejko, Z., Curtis, C. E., & Ma, W. J. (2018). Strategic allocation of working memory resource. *Scientific Reports*, *8*, 16162. https://doi.org/10.1038/s41598-018-34282-1

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. https://doi.org/10.1038/nature06860

Zhao, F., Zeng, Y., Wang, G., Bai, J., & Xu, B. (2018). A Brain-Inspired Decision Making Model Based on Top-Down Biasing of Prefrontal Cortex to Basal Ganglia and Its Application in Autonomous UAV Explorations. *Cognitive Computation*, *10*(2), 296–306. https://doi.org/10.1007/s12559-017-9511-3

Zipser, D. (1991). Recurrent Network Model of the Neural Mechanism of Short-Term Active Memory. *Neural Computation*, *3*(2), 179–193. https://doi.org/10.1162/neco.1991.3.2.179