frontiers | Frontiers in Computational Neuroscience

TYPE Review
PUBLISHED 21 December 2022
DOI 10.3389/fncom.2022.1060101

OPEN ACCESS

EDITED BY
Chang-Eop Kim,
Gachon University, South Korea

REVIEWED BY
Shinsuke Suzuki,
The University of Melbourne, Australia
Sang-Woong Lee,
Gachon University, South Korea
Hava T. Siegelmann,
Rutgers, The State University of New Jersey, United States

*CORRESPONDENCE
Sang Wan Lee
sangwan@kaist.ac.kr

RECEIVED 02 October 2022
ACCEPTED 30 November 2022
PUBLISHED 21 December 2022

CITATION
Lee JH, Leibo JZ, An SJ and Lee SW (2022) Importance of prefrontal meta control in human-like reinforcement learning.
Front. Comput. Neurosci. 16:1060101. doi: 10.3389/fncom.2022.1060101

Importance of prefrontal meta control in human-like reinforcement learning

Jee Hang Lee[1], Joel Z. Leibo[2], Su Jin An[3] and Sang Wan Lee[3,4,5,6,7]*

[1]Department of Human-Centered Artificial Intelligence, Sangmyung University, Seoul, South Korea, [2]Google DeepMind, London, United Kingdom, [3]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, [4]Program of Brain and Cognitive Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, [5]KAIST Center for Neuroscience-Inspired Artificial Intelligence, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, [6]KAIST Institute for Health Science and Technology, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, [7]KAIST Institute for Artificial Intelligence, Korea Advanced Institute of Science and Technology, Daejeon, South Korea

Recent investigation on reinforcement learning (RL) has demonstrated considerable flexibility in dealing with various problems. However, such models often experience difficulty learning seemingly easy tasks for humans. To reconcile the discrepancy, our paper is focused on the computational benefits of the brain's RL. We examine the brain's ability to combine complementary learning strategies to resolve the trade-off between prediction performance, computational costs, and time constraints. The complex need for task performance created by a volatile and/or multi-agent environment motivates the brain to continually explore an ideal combination of multiple strategies, called meta-control. Understanding these functions would allow us to build human-aligned RL models.

KEYWORDS

reinforcement learning, neuroscience of RL, prefrontal meta control, model-based and model-free RL, human-aligned RL models

1. Introduction

Suppose a new game was released. Reading the detailed "how to play" instructions would be a desirable first step toward mastering the game. However, when such instruction is not offered, we might consider playing the game without prior knowledge. This undirected approach often works in practice. In fact, it is often the case that we can understand the gist of the game after only a few trials, including its rules, goals, and environmental structure, and how to draw up a draft strategy before deliberate planning. Subsequent experience brings further improvement from such initial information, rapidly refining a strategy suitable to various contexts and goals.

This example briefly describes the human capability of forming optimal behaviors based on learning from experiences, the so-called theory of reinforcement learning (RL) in the human mind. Inspired by the psychology and neuroscience of such human behavior, RL theory provided the mathematical scaffolding to describe how humans learn from past experiences (Schultz, 1998; Sutton and Barto, 1998). Due to breakthroughs in deep learning and a steep increase in computing power, RL theory and practice have led to remarkable advances in our ability to design artificial agents with super-human

performance. Results demonstrate the applicability of RL in various domains, including games (Mnih et al., 2015; Silver et al., 2016, 2017a,b, 2018; Schrittwieser et al., 2020), large-scale Markov decision problems (Sutton and Barto, 1998; Szepesvári, 2010; Sigaud and Buffet, 2013; Covington et al., 2016; Evans and Gao, 2016; Dulac-Arnold et al., 2021), and non-linear and stochastic optimal control problems without explicit representations of environments (Bertsekas and Tsitsiklis, 1995; Si, 2004; Busoniu et al., 2010; Fujimoto et al., 2019; Vecerik et al., 2019).

Nevertheless, the computational principles underpinning RL algorithms still differ from the manner in which human RL works. The goal of RL is to develop a policy that specifies the choice of action for each state of the world so as to maximize the expected amount of future reward (Sutton and Barto, 1998). It is common to divide the space of principles supporting RL algorithms into two categories called, respectively, model-free (MF) and model-based (MB). Most state-of-the-art RL agents only incorporate the model-free (MF) RL principle (Doya et al., 2002; Daw et al., 2005), whereas human brains employ both MF and MB principles simultaneously (Dolan and Dayan, 2013).

In animals, MF RL is guided by the dopaminergic striatal system (Montague et al., 1996; Schultz et al., 1997). Based on the "trial-and-error" concept, MF RL incrementally updates the values of actions based on reward prediction error, which is the quantity that represents the discrepancy between the agent's prediction regarding reward and the actual rewards it receives from the environment. Iterating this process is a way to improve a policy, continually adjusting it to obtain more rewards until convergence. Many repetitions of a contingency are usually required to incorporate it into a policy by MF RL. As a result, policies learned by MF RL are said to be habit-like and automatic, making it harder to respond quickly to changing context. For instance, when there is a sudden change in the environment, or when a new goal is established, a major number of classical MF RL agents is likely to require re-learning to adapt to the change. This is often time-consuming and computationally inefficient. In particular, it requires a massive amount of new experience, since the effect of old information will only slowly decay by averaging as more and more contradictory information comes in later.

Humans can learn to adapt to environmental changes with a small amount (or an almost absence) of experience. As illustrated by the game example, human brains can go beyond the MF RL strategy to achieve impressive scores with high efficiency, speed, and flexibility in learning and behavior control—whether the game is entirely new or not (Lake et al., 2017). Recent progress in decision neuroscience indicates that the human brain also employs another learning strategy besides MF RL, called model-based (MB) RL (Daw et al., 2005; Dolan and Dayan, 2013). Using this strategy, the brain learns about different structures in the world, such as a state space or reward structure, leading to a deliberate behavioral policy that is sensitive to changes in the structure of the state space or goals (Kuvayev and Sutton, 1996; Doya et al., 2002; Daw et al., 2005; O'Doherty et al., 2015). In addition, humans have the ability to learn from a small number of observations: neural computations underlying rapid learning comprising so-called one-shot inference, can be distinguished from those used for incremental learning (Daw et al., 2006; Boorman et al., 2009; Badre et al., 2012; Lee et al., 2015; Meyniel et al., 2015). One-shot inference refers to the situation in which an agent learns rapidly from only a single pairing of a stimulus and a consequence, as often required for goal-driven choices. Incremental learning refers to the situation in which an agent gradually acquires new knowledge through "trial and error," as seen in MF RL. A proper combination of these two types of learning may guide optimal behavior control in various situations within a relatively short time with access to a relatively small set of experiences (or data) in practice.

A neural mechanism called "meta-control" on top of multiple systems utilized for human RL accounts for behavioral flexibility, memory efficiency, and rapid learning speed in humans. Recent neuroimaging studies and the computational modeling used in RL studies have identified not only the respective neural correlates of MF and MB RL (Gläscher et al., 2010), but also the neuro-anatomical circuits responsible for arbitration between the two types of RL. The above studies are based upon the proposal that the arbitration is governed by the relative amount of uncertainty in the estimates of the two systems (Lee et al., 2014). Another meta-control ability is that of determining when to learn incrementally or rapidly to make inferences regarding the state of the world (learning from a few observations). This meta-control process can be extremely useful in guiding behavior during learning (Meyniel et al., 2015), in deciding whether to explore a new alternative or to pursue the currently available option (Daw et al., 2006; Badre et al., 2012), and in evaluating an alternative course of action (Boorman et al., 2009). It is noted that learning regarding the state-space is necessary for MB RL because the probabilistic representation of the state-space is an essential component required for computing the expected amount of future reward.

Understanding how the human brain implements these abilities, which state-of-the-art RL algorithms do not possess, would help us improve the design of RL algorithms as follows. First, an agent with an MB RL strategy would learn about the model of the environment and leverage this knowledge to guide goal-driven behavior. This includes action planning to foresee future episodes, even if they are computationally expensive, based on the model of the environment that the RL agent has in mind. Second, we expect that an RL agent with a rapid inference ability would learn a model of its environment from a very small number of samples, expediting the MB RL process. Finally, adaptive control of these functions serving RL, dubbed meta-control of RL, would resolve the trade-offs among prediction performance, computational load, and training efficiency.

This paper is organized as follows. In Section 2, we briefly overview the computational principles of RL and explore a few situations posing significant challenges to recent RL algorithms. We then discuss how the brain solves these challenges in a point-by-point manner in Section 3. We will specifically discuss MB and MF RL, one-shot inference in MB RL, and the meta-control process over these strategies. This discussion leads us to potential research questions presented in Section 4. A concluding remark is provided in Section 5.

2. Reinforcement learning—Basic ideas and challenges

2.1. Basic concepts

The theory of RL is a normative framework to account for the general principle describing how value-based, sequential decision-making takes place in humans (Mnih et al., 2015). RL algorithms in computer science are usually based on Markov Decision Processes (MDPs) (Bellman, 1957), which commonly model various sequential decision problems incorporating uncertainty in the environment.

Sequential choices, which occur in a range of real-world problems, is a fundamental task that any intelligent agents (including humans and animals) encounter in extended actions/interactions with their environment (Littman, 1996). In this circumstance, the agents iteratively try to make an optimal decision to achieve a goal in a sequential manner, through learning and inference. The agents need to act on what they have learned, use them to infer the decision which can possibly bring about the best outcomes, and learns from the obtained outcome for decision-making in the future. With this aim in mind, agents are capable of dealing with these sequential decision problems by means of *programming*, *search and planning*, or *learning* approaches. In general, agents who learn to make optimal decisions take into consideration a combinatorial approach—they carry out the *planning* in order to establish long-term actions in uncertain domains on the foundation of the *learning* about the environment. Sometimes the choice between exploiting what they already know and exploring new options that may lead to better outcomes (or worse) takes place for the purpose that either maximizing the effect of actions or toward a higher learning performance assuring a better model of an environment enabling the better outcomes (van Otterlo and Wiering, 2012).

These sequential decision problems are usually solved either by learning and planning given a model of the MDP referred to as MB RL, or by learning through actions/interaction with an unknown MDP referred to as MF RL. During the process, the desirability or undesirability of actions that agents choose in each state, and their effects are evaluated by a reward codified in a single scalar objective function. The objective of the agents is then the maximization of the (discounted) expected sum of the scalar reward at each step over time (Roijers et al., 2013).

A solution to the MDP is characterized by the Bellman optimality equation (Sutton and Barto, 1998).

$$\begin{aligned} Q^*(s,a) &= E_{(s,a,s')}\Big[R + \gamma \max_{a'} Q^*(s',a')\Big] \\ &= \sum_{s'} P(s,a,s')\Big(R + \gamma \max_{a'} Q^*(s',a')\Big) \end{aligned} \tag{1}$$

where the tuple $\langle s, a, s' \rangle$ refers to the current state s, an action a, and the state in the next time step s', and $Q(s,a)$ refers to the state-action value. $P(s,a,s')$ and R refer to the state-action-state transition probability and an immediate reward, respectively. It specifies that the value estimate for states and actions is based on the expectation over a state-space distribution of the quantity consisting of the amount of immediate reward plus the value estimate of the possible next state.

The goal of RL is to learn an optimal policy by estimating the expected amount of reward for each state or action $Q^*(s,a)$. Classical RL agents have employed various iterative methods (Sutton, 1988; Watkins, 1989; Barto and Duff, 1994; Singh and Sutton, 1996), but learning exact representations of value functions in high dimensional state space is often computationally intractable. In recent deep reinforcement learning research, non-linear, parameterized function approximation techniques are used to represent value functions, policies, and models of the environment. The combination of RL with deep learning has led to rapid advances in RL algorithm design with outstanding performance in many applications, including games, robot control and simulated environments (Silver et al., 2014; Lillicrap et al., 2015; Mnih et al., 2015, 2016; Van Hasselt et al., 2016; Kalashnikov et al., 2018; OpenAI, 2018; Vecerik et al., 2019). Mounting evidence suggests that similar algorithms are present in the mammalian brain and are embedded in different types of human decision-making systems (Daw et al., 2005; Dayan and Daw, 2008; Rangel et al., 2008; Balleine and O'doherty, 2010; Dolan and Dayan, 2013; Gesiarz and Crockett, 2015).

2.2. Major challenges

The combination of deep learning and RL used in the state-of-the-art RL algorithms has shown dramatic success in both theory and practice. Nonetheless, the computational principle of deep RL is still different from the way human RL works. Let us recall Bellman's optimality equation (Equation 1). The optimality of the policy is in principle determined by the expected amount of *long-term cumulative rewards* over a *state-action-state transition probability distribution*, each of which we call *rewards* and the *model* of the environment, respectively.

A majority of RL agents, including state-of-the-art RL algorithms, usually incorporate the MF RL principle. Here, the

state-action-state transition probability is often replaced with an empirical sampling from the environment; it does not require an explicit representation of the model of the environment. Recent works such as *Muesli* (Hessel et al., 2021) exhibited the capacity to learn a model of sophisticated state representation as opposed to this, but they are mostly limited to show the planning capacity on top of the model learned i.e., learning from a simulation of possible futures using a model of an environment. This is likely to fail to demonstrate the human's capability to introspect their thought process (MB RL), and to account for the human's behavioral flexibility in arbitrating between MF and MB, the characteristics called "meta-control."

It thus lacks the ability to develop goal-directed policies, making itself less flexible although RL algorithms are a simple and effective way to find and explore better policies. For example, a major number of classical MF RL agents requires all new learning when a new goal, such as "find a piece of cheese instead of a cup of water," "achieve the lowest possible score," or "achieve a goal without embarrassing your opponent," is established (Lake et al., 2017). This is time-consuming and inefficient; it requires a lot of experience (resampling from the environment) because an MF RL agent mostly is likely to rely on the retrospective learning principle i.e., learning from past experience. This in consequence makes the training of RL agents slower and less flexible, which has been a challenge from a computational point of view.

Due to the perceived shortcomings of MF RL approaches, a growing number of MB RL algorithms have been suggested (Moerland et al., 2020) as neuroscientific findings on MB RL agents have been shown to achieve goal-directed behavioral adaptations (Doya et al., 2002; Lee et al., 2014). MF RL algorithms require a massive amount of experience to learn. This in turn leads to a significant diminution in their ability to rapidly adapt to dynamic environments where a context and its associated required tasks are frequently changed. As widely known, MB RL algorithms appear to have many potential gains to this end, such as sample efficiency, or fast adaptation to environmental changes (Daw et al., 2011; Moerland et al., 2020).

However, it is not entirely clear that MB RL is always superior to MF RL in sample efficiency, particularly in a single task. It is still in doubt that the time for learning a model and planning with the model is quicker than that for learning an optimal policy directly from the episodes under this circumstance. In addition, it is arguable whether MB RL is always better than MF RL with respect to its fast adaptation ability (Kim and Lee, 2022; Wan et al., 2022). For instance, a recent MF RL algorithm was able to achieve zero-shot learning to new goals that it never experienced during learning (Stooke et al., 2021). This algorithm was clearly MF RL, since it does not possess any model learning or planning capacity.

There is mounting evidence in decision neuroscience that has led to clarification of the principles of how the brain solves the aforementioned issues. One line of evidence suggests that the human brain employs not only MF RL but also MB RL. Other evidence indicates that the brain has the ability to learn from a few or even a single observation(s) in a process dubbed "one-shot inference" (Lee et al., 2015; Garcia and Bruna, 2018). Specifically, the human brain is engaged in determining when to learn incrementally or rapidly to make inferences regarding the state of the world. This process can be extremely useful in guiding an agent's behavior during learning (Meyniel et al., 2015), in deciding whether to explore a new alternative or pursue a currently available option (Daw et al., 2006; Badre et al., 2012), or in evaluating an alternative course of action (Boorman et al., 2009). In the following section, we will investigate how the human brain implements MF/MB RL itself and one-shot inference to guide MB RL, and how these different functional units are controlled in the brain.

3. Computational principles of RL in the human brain

It is widely accepted that human behavior is accounted for by two different behavior control strategies: stimulus-driven and goal-directed behavior control (for a more extensive review on these strategies, see O'Doherty et al., 2017). Historically, the brain has been thought to exert stimulus-driven behavior control (Thorndike, 1898). According to this theory, a biological agent exhibits habitual response patterns that are highly insensitive to changes in the consequences of its actions (Thibodeau et al., 1992). This has been contrasted with the idea of goal-directed behavior control, wherein deliberative actions are motivated by a specific goal (Tolman, 1948; Valentin et al., 2007).

Each strategy provides a different complementary solution considering accuracy, speed, and cognitive load (O'Doherty et al., 2017). Goal-directed behavior control allows humans to pursue adaptation to environmental changes without re-experiencing (or re-sampling) (Tolman, 1948). However, it is cognitively demanding and therefore slow. In contrast, stimulus-driven behavior control is cognitively productive, automatic, and fast despite being fragile in a volatile environment (O'Doherty et al., 2017). It appears that humans use specific principles to determine the dominating type of control to guide behavior in different contexts (Dickinson et al., 1983).

The above behavioral findings highlighting the two contrasting behavior control strategies beg the question of whether and how the human brain implements respective RL strategies (Doya, 1999; Daw et al., 2005; O'Doherty et al., 2015). As Daw et al. (2005) have proposed, the two distinct types of RL (MF and MB RL) guide human behavior, and can account for habitual and goal-directed behavior control, respectively. In the following section, we will focus on exploring the neural correlates of MB and MF RL in order to better understand the computational principles underlying RL in humans.

3.1. Neural correlates of RL

Animals learn to survive by making choices that lead to the receipt of rewards (e.g., food or water) and avoidance of penalties (e.g., sickness or death). In doing so, the animal should be able to estimate the value of each environmental option. This ecological conception has motivated research on the neural representations of value signals in the brain (Camerer et al., 2005; Padoa-Schioppa and Assad, 2006; Glimcher and Fehr, 2013; Juechems et al., 2017). Such investigations indicate that the value signals are found in several brain regions including the amygdala, orbitofrontal cortex, ventromedial prefrontal cortex, and ventral and dorsal striatum (Saez et al., 2015; O'Doherty et al., 2017), as well as the parietal and supplementary motor cortices (Hampton et al., 2006; Gläscher et al., 2008; Boorman et al., 2009).

The quantity representing the discrepancy between predicted future rewards and actual rewards, called a reward prediction error (RPE) signal, is required to update the value signal. The reward prediction error signal encodes the phasic activity of dopamine neurons, as seen in Figure 1A (Schultz et al., 1997).

Since the reward prediction error plays a key role in RL, the RL framework has been used in a wide range of neuroscience disciplines. In particular, RL has been used to explain the computational functions of neuromodulators such as dopamine, acetylcholine, and serotonin (Sutton, 1988; Sutton and Barto, 1998). Phasic firing patterns of dopaminergic neurons reflect the characteristics of temporal difference prediction error in humans (Niv, 2009). Earlier studies have reported that dopaminergic neurons convey information regarding current events and the predictive value of the current state, and that the circuitry involving dopaminergic nuclei uses this information to compute a temporal difference-style reward prediction error (Christoph et al., 1986; Floresco et al., 2003; Nakahara et al., 2004; Geisler and Zahm, 2005; Matsumoto and Hikosaka, 2007). Human neuroimaging studies have also reported evidence for the presence of temporal difference prediction error signals in dopamine neurons (Glimcher, 2011; Lee et al., 2012) and reward/state prediction errors in the human brain, as shown in Figure 1B (Lee et al., 2014). In summary, these findings indicate that MF RL, including the temporal difference model, is by far the most appropriate theoretical principle to explain how animals, including humans, learn to survive.

3.2. Trade-off between prediction performance and computational costs: Model-based and model-free RL

Typical MF RL algorithms can successfully account for choice patterns in simple decision-making tasks. However, they fail to explain the choice patterns in multi-stage Markov decision tasks (Gläscher et al., 2010; Lee et al., 2014). There is thus a need to test the hypothesis that additional type(s) of RL strategies may be used at different time points (Dickinson et al., 1983).

Behavioral evidence indicates that there are at least two types of behavior manifesting at different time points. For example, in the early stage of training, an agent is likely to select an action based on predicted outcomes, while later on, action is elicited by a prior antecedent stimulus. The former and latter are called goal-directed and habitual behavior, respectively. Accumulating evidence supports the existence of separate neural substrates guiding these two types of behavior (Dickinson et al., 1983; Dayan and Berridge, 2014; Nasser et al., 2015; O'Doherty et al., 2017).

Figure 2A provides an example of the above phenomena in the context of the strategic game Tic-Tac-Toe. An MF RL agent would attempt to choose the next strategy so as to win the game and is in favor of maximizing the value. Here, the MF RL (right panel in Figure 2A) would update the value *via* sampling without consideration of the model of the game. Such action patterns associated with habitual behavior are accounted for by MF RL algorithms, which learn the values of actions based on reward prediction errors in a process of backward learning. In contrast, an MB RL agent (middle panel in Figure 2A) would first learn a model for the game (i.e., state-action-state transition probability). It would then decide on an option to win the game. The action patterns associated with this goal-directed behavior are accounted for by MB RL, which uses state-prediction errors to learn the values of actions online by combining information regarding the estimated outcome and the learned model of the environment. Therefore, an exploration stage to model an environment (e.g., a few more plays with the opponent in this case) is necessary in this context.

The major distinction between MB and MF RL is the assumption that the agent uses the knowledge of the environment to update action values as described above. For example, the MB learner computes the expected future outcome using a state-action-state transition probability distribution, whereas the MF learner does not rely on the availability of a perfect state-transition model. Neural evidence supports this assumption, as shown in Figures 2A, B. Based on the prediction error signals described in Figure 1B, a neural mechanism would process value signals from several brain regions in human RL. These brain regions include the dorsomedial prefrontal cortex, which encodes an MB value (Wunderlich et al., 2012; Doll et al., 2015), the posterior putamen, which encodes an MF value (Tricomi et al., 2009), and the ventromedial prefrontal cortex, which integrates MB and MF values (Boorman et al., 2009; Hare et al., 2009; Rushworth et al., 2012).

This computational distinction suggests that there is an inevitable compromise between the two strategies. MB RL provides more accurate predictions than MF RL in general, though both processes converge upon an optimal behavior

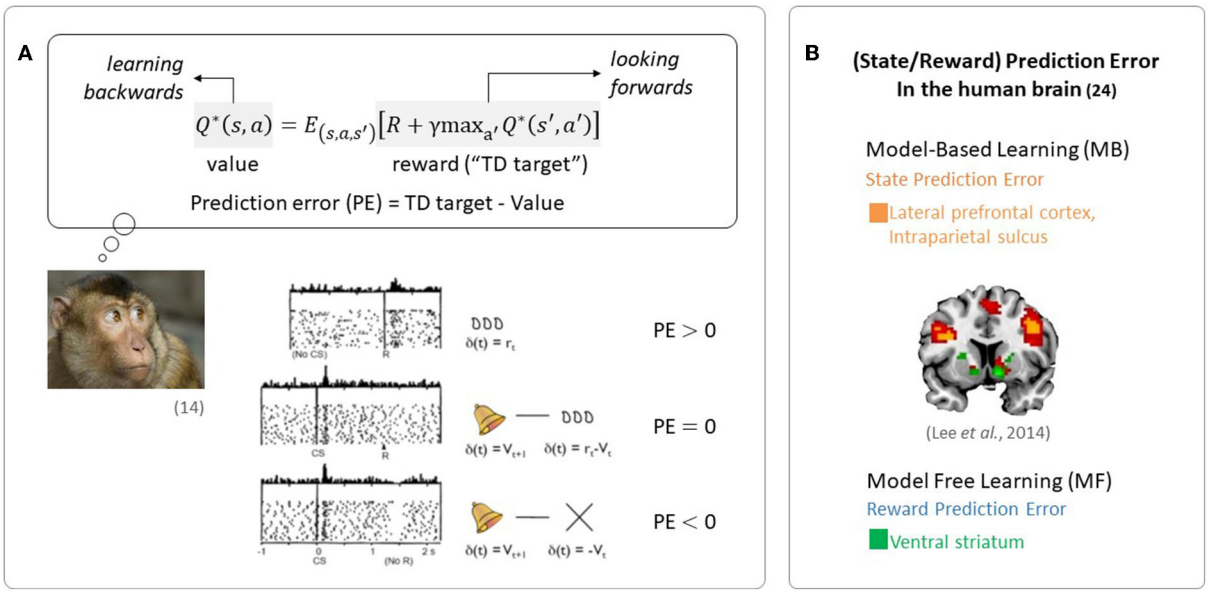

FIGURE 1
Evidence of RL in the brain. **(A)** Midbrain dopamine neurons encode information regarding the discrepancy between a predicted outcome and the actual outcome, an essential component in regulating RL (Figures are adapted from Schultz et al., 1997). **(B)** Neural evidence for the use of prediction errors in the human brain: state prediction error and reward prediction error (the figure was taken from Lee et al., 2014).

strategy. As a result, performance differences between the two strategies diminish over time. Nevertheless, MB RL is computationally heavier than its counterpart. This indicates that there is a trade-off between prediction performance and computational costs.

3.3. Trade-off between prediction performance and time constraints: Incremental and one-shot learning

It is not surprising that RL agents require a sufficient number of experiences to fully learn causal relationships in the presence of different environmental factors. This is the basic principle underlying incremental inference. In this case, the agents gradually learn through trial and error to identify stimuli leading to particular consequences. There has been substantial progress in understanding the computational mechanism underlying incremental inference. Various algorithms, such as the Rescorla-Wagner rule (Rescorla and Wagner, 1972), the probabilistic contrast model (Jenkins and Ward, 1965), the associative learning model (Pearce and Hall, 1980; McLaren and Mackintosh, 2000), and Bayesian causal inference (Griffiths and Tenenbaum, 2009; Holyoak et al., 2010; Carroll et al., 2011) provide computational accounts for the behavioral characteristics associated with incremental inference. Note that an MB RL agent would gradually learn about the model of the environment if the incremental inference strategy is used.

Unlike in incremental inference, the agent sometimes learns the associations very rapidly after a single exhibition of a novel event never experienced before. This is called "one-shot" inference. This ability has been demonstrated in animal learning (Moore and Sellen, 2006; Schippers and Van Lange, 2006; Garety et al., 2011; Moutoussis et al., 2011) and object categorization (Fei-Fei et al., 2006). Although the distinctive case of one-shot inference has been relatively well-discussed in behavioral studies (Moore and Sellen, 2006; Garety et al., 2011; Moutoussis et al., 2011), its computational mechanism has received scant attention. Lee et al. (2015) investigated the computational and neural mechanisms underlying one-shot inference. They presented evidence indicating that the level of uncertainty regarding "cause-effect" relationships mediates the transition between incremental and one-shot inference. For example, more causal uncertainty leads to the assignment of a higher learning rate to a stimulus. This in turn helps resolve uncertainty and facilitates very rapid one-shot inference. This explains when and how one-shot inference occurs in preference to incremental inference, and how the brain is able to switch between the two learning strategies.

Figure 2 provides an example of the behavioral and neural evidence regarding incremental and one-shot learning in MB RL. Incremental inference (bottom-left panel in Figure 2A) usually requires considerable experience or frequent exposure to a cause-effect pairing to learn the causal relationship between the two events. On the other hand, one-shot inference learning (top-left panel in Figure 2A) requires only a single exposure to

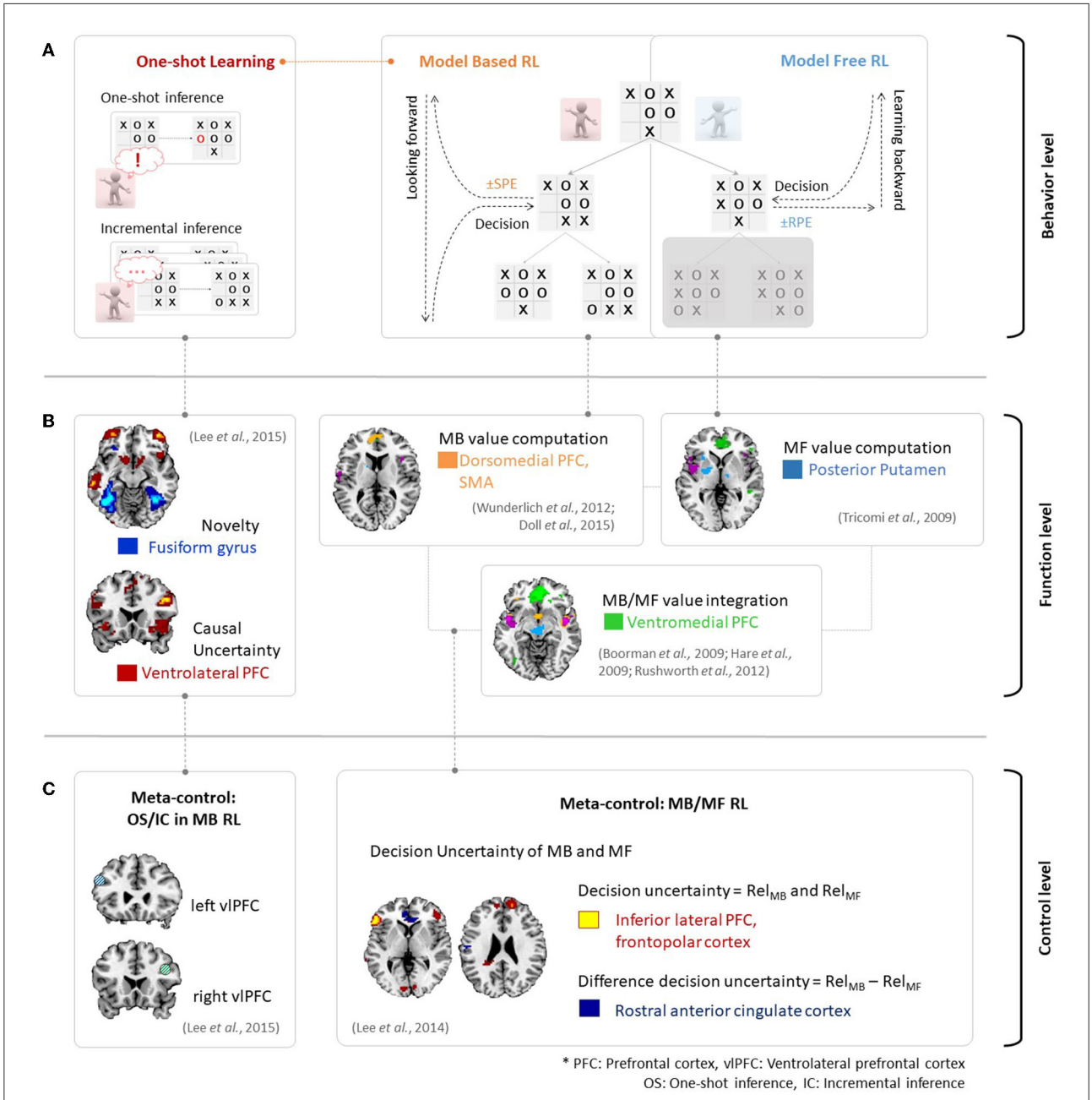

FIGURE 2
(A) Behavioral evidence for MB/MF RL and incremental/one-shot inference in MB RL in the context of the game of Tic-Tac-Toe. **(B)** Neural evidence for MB/MF RL and one-shot inference. Figures are adapted from Wunderlich et al. (2012); Doll et al. (2015) for MB value, Tricomi et al. (2009) for MF value, Boorman et al. (2009); Hare et al. (2009); Rushworth et al. (2012) for MB/MF value integration and Lee et al. (2015) for one-shot inference. **(C)** Neural evidence of meta-control over MB and MF RL (right figure is adapted from Lee et al., 2014), and incremental and oneshot-inference (left figure is adapted from Lee et al., 2015).

a cause-effect pairing or a single experience. In the Tic-Tac-Toe game context, a player may establish the winning strategy based on the number of game plays when he or she is using incremental inference. However, a player may also establish the winning strategy based on a single novel experience when using one-shot inference. Recent findings proposed by Lee et al. (2015) describe the neural activity associated with one-shot learning. As seen in Figure 2B, the ventrolateral prefrontal cortex (vlPFC) encodes causal uncertainty signals and the fusiform gyrus encodes the novelty of a given cause-effect pair. The fusiform gyrus then plays a crucial role in the implementation of switching control between incremental and one-shot inference.

Note that there is a trade-off issue here as well. For instance, an RL agent that learns incrementally provides more reliable predictions but is slower than one based on one-shot inference. Determining the strategy that the agent pursues would depend on prediction performance and time constraints.

3.4. Prefrontal meta-control to resolve the performance-efficiency-speed trade-off

As described earlier, the brain exerts control over behavior using multiple complementary strategies: (i) MB and MF RL, and (ii) incremental and one-shot learning. The former addresses the trade-off between prediction performance and computational efficiency, while the latter addressed the trade-off between prediction performance and learning speed.

These findings beg the question of whether the brain implements a principled policy to arbitrate between the two sets of strategies. Earlier theoretical work hypothesized that there exists a brain region that determines the amount of influence of each strategy on behavior (Daw et al., 2005). Subsequent studies have found evidence for the existence of such a mechanism in the human brain: the arbitration between MB and MF RL (Lee et al., 2014), and that between incremental and one-shot learning (Lee et al., 2015).

Such arbitration processes are predominantly found in the ventrolateral prefrontal cortex (Lee et al., 2014, 2015), as seen in Figure 2C. On the one hand, the ventrolateral prefrontal cortex computes the decision uncertainty in MB and MF RL while taking into account the prediction error (reward prediction error in MF RL and state prediction error in MB RL). This results in the model choice probability (P_{MB}). The ventrolateral prefrontal cortex chooses the more reliable system (either MF or MB) depending on P_{MB}. This would in turn control the behavior, as appropriate given the situation (right panel in Figure 2C). On the other hand, the ventrolateral prefrontal cortex is also involved in the choice of the mode of inference depending upon the degree of functional coupling between the ventrolateral prefrontal cortex and hippocampus. Specifically, the degree of functional coupling increases when one-shot inference is predicted to occur and decreases when incremental inference is predicted in the hippocampus.

The main finding of the above studies is that the key variable for arbitration is uncertainty in the prediction performance for each strategy. For example, let us assume that the MF RL agent has recently indicated high reward prediction errors, while the MB RL system has indicated low state prediction errors simultaneously at a particular moment. This would imply that the MF agent is less reliable while the MB system (i.e., goal-directed system) is warranted at this moment. In this situation, the behavioral policy would be influenced by the MB system while the brain reduces the influence of the MF system. However, it is noted that MB RL is a computationally expensive process, so the brain seems to resort to the MF RL in situations wherein the agent does not gain considerable benefit from learning the environment, such as when the environment is sufficiently stable for MF RL to make precise predictions, or when it is extremely unstable to the extent that the predictions of MB RL become less reliable than those of MF RL.

The same principle applies to the arbitration between incremental and one-shot learning. When the uncertainty in the estimated cause-effect relationships is high, the brain tends to transition to one-shot learning by increasing its learning rate. This would help the agent quickly resolve uncertainty in predicting outcomes. However, when the agent is equally uncertain about all possible causal relationships, the brain seems to resort to incremental learning. In summary, one of the important goals of RL in the human brain is to reduce the total amount of uncertainty in prediction performance. In doing so, it naturally resolves the trade-offs among prediction performance, computational efficiency, and learning speed. When using the above approach (based on the accuracy of predictions), the cognitive effort required for behavior control (FitzGerald et al., 2014) and the potential cumulative benefits inferred by an MB strategy (Pezzulo et al., 2013; Shenhav et al., 2013) are often taken into account in the meta-control of MB and MF RL.

3.5. Multi-agent model-based and model-free RL

New issues arise when you have a system consisting of multiple interacting agents. There are two basic cases, the cooperative case where agents have fully aligned objectives. This case is important for many applications (Claus and Boutilier, 1998; Panait and Luke, 2005). The other case, more common in nature, is called non-cooperative. Interactions in non-cooperative situations may be either fully competitive (zero-sum in the language of game theory) or partially competitive.

AlphaGo (Silver et al., 2016), an agent that defeats top human Go players is arguably the most successful example to date of an artificial system that combines MB and MF mechanisms. However, it does not attempt to combine them in a biologically-plausible manner. AlphaZero works by alternating learning a policy network by MF RL and improving it by Monte-Carlo tree search (an MB RL method) (Silver et al., 2017a,b). One reason MB methods work so well on board games but not in other domains is that a perfect model is available in these cases. The rules of the game are a complete description of the one-step transition function, and they are assumed to be known *a priori* and perfectly. This assumption is true of board games like Go and Chess but it does not even hold for games

of imperfect information like Poker, much less for complex real-world environments.

Many multi-agent interactions that have been important in human evolution have partially competitive and partially aligned incentives. In particular, there are social dilemma situations. These are situations where individuals profit from acting selfishly, but the group as a whole would do better if all individuals curbed their egoism and instead acted toward the common good (Kollock, 1998). Famous social dilemmas arise in cases where there are resources that have properties making it difficult for any individual to exclude others from accessing them. For example, if all community members may access a common fishery, each individual is expected to catch as many fish as they can since they each gain from every additional fish they catch. But if all behave this way then the stock of still uncaught fish will be depleted too quickly causing the fishery to decline in productivity. This scenario and others like it have been called the tragedy of the commons (Hardin, 1968). Diverse theories of human evolution agree that navigating social dilemmas like these have been critical, especially as we have become more and more of an obligate cultural species, unable to survive even in our own ancestral ecological niche (hunting and gathering) without significant cooperation (Henrich, 2015).

There is a large classical literature concerned with agents that cooperate in abstracted matrix game models of social dilemma situations like iterated prisoner's dilemma (Rapoport et al., 1965; Axelrod and Hamilton, 1981). More recently, several algorithms have been described that achieve cooperation in more complex temporally extended settings called Markov games. Formally this setting is a straightforward generalization of Markov decision processes to multiple players (Littman, 1994). Some recent algorithms can be seen as MB (Kleiman-Weiner et al., 2016; Lerer and Peysakhovich, 2017), like in AlphaGo, the agent is assumed to have a perfect model of the rules of the game. These algorithms work in two stages, first there is a "planning" stage where the agent simulates a large number of games with itself and learns separate cooperation and defection policies from them by applying standard MF RL methods toward both selfish and cooperative objectives independently. Then in the execution phase, a tit-for-tat policy is constructed and applied using the previously learned cooperate and defect policies.

Some recent algorithms have sought to break down the strict separation between planning and execution stages and instead work in a fully on-line manner. One model-based example is the LOLA algorithm (Foerster et al., 2018). In addition to assuming perfect knowledge of the game rules, this model also assumes that agents can differentiate through one another's learning process. That is, it assumes that all agents implement a policy gradient learning algorithm. This allows agents to "learn to teach" since they can isolate the effects of their actions on the learning of others. It is possible that learned models for the environment and the learning updates of other players could be substituted in the process, but this has not yet been shown convincingly.

Another line of research on resolving multi-agent social dilemmas is based on MF RL. It drops the need for assuming perfect knowledge of the game rules (a perfect model) and works most naturally in the standard fully online setting (Leibo et al., 2017; Perolat et al., 2017). Considerable evidence from behavioral economics shows that humans have inequity-averse social preferences (Fehr and Schmidt, 1999; Henrich et al., 2001; McAuliffe et al., 2017). One algorithm in this class, proposed first for matrix games (Gintis, 2000; De Jong et al., 2008), and later extended to Markox games (Hughes et al., 2018), modifies standard MF RL to use the following inequity-averse reward function.

Let $r_1, \ldots, r_N$ be the payoffs achieved by each of N players. Each agent receives the subjective reward

$$\begin{aligned} U_i(r_i, \ldots r_N) = \; & r_i \\ & - \frac{\alpha_i}{N-1} \sum_{j \neq i} \max\left(r_j - r_i, 0\right) \\ & - \frac{\beta_i}{N-1} \sum_{j \neq i} \max\left(r_i - r_j, 0\right), \end{aligned} \tag{2}$$

The alpha parameter controls disadvantageous inequity aversion ("envy") and the beta parameter controls advantageous inequity aversion ("guilt"). Simulations of agents with high beta parameters show that they are able to discover cooperative equilibria more easily than selfish agents since individuals are disincentivized from improving their policy in directions from which they benefit at the expense of the rest of the group. In addition, agents with high alpha parameters sometimes appear to act as "police", punishing anti-social behavior in other agents, thereby disincentivizing defection and promoting cooperative outcomes (Hughes et al., 2018).

Disadvantageous inequity aversion is thought to be present in other species while advantageous inequity aversion may be uniquely human (McAuliffe et al., 2017). Both depend on the same neural circuity for valuation that support non-social decision-making (Fehr and Camerer, 2007). One especially relevant study found that activity in the ventral striatum and ventromedial prefrontal cortex were significantly affected by both advantageous and disadvantageous inequity (Tricomi et al., 2010).

4. Potential research directions

Here, we show that the convergence of computer science and decision neuroscience can extend our understanding of how the human brain implements RL. Given the evidence thus far, human RL appears to utilize not only multiple systems, but also a flexible meta-control mechanism to select among them.

Figure 3C is a schematic diagram summarizing the brain network for human RL. The multiple systems facilitating human RL comprise (i) an MB system that is flexible but cognitively demanding, (ii) an MF system that is simple but inflexible, (iii) an incremental inference system that is careful while learning but slow, and (iv) a one-shot inference system that is fast in learning but has the potential to misattribute (Figures 4A, B). Based on the characteristics of the multiple systems, human RL flexibly chooses the most appropriate system while taking into account performance, efficiency, and speed. When situated in a completely new environment, meta-control accentuates speed and performance: the human RL system learns about the environment as fast as possible using the one-shot inference system while utilizing the cognitively demanding MB system to maximize performance with (relatively) lower confidence based on prior knowledge. In other situations, meta-control prioritizes efficiency and speed: the human RL system uses simple and efficient MF systems to maximize performance with accurate knowledge that is carefully constructed over time using the incremental inference system (Figure 4C).

This principle of human RL could shed light on the manner in which fundamental issues in engineering are resolved with a focus on the trade-offs among performance, efficiency, and speed, particularly in the design of artificial agents and their embodiments, robots. As highlighted in the game example, the discordance between human RL and algorithmic RL is seen in Figures 3B, C lies in the existence of the ability to flexibly control the agent's behavior in the face of dynamic changes in the environment (e.g., goals and rewards). Therefore, the principle of human RL will fuel the advent of embodied algorithms enabling RL agents to show super-human or super-artificial intelligence performance.

Of course, previous studies focusing on individual RL systems have substantially contributed to the birth of various RL algorithms. As seen in Figure 3B, MF RL algorithms, such as Monte-Carlo methods (Barto and Duff, 1994; Singh and Sutton, 1996), TD methods (Sutton, 1988), Q-Learning (Watkins and Dayan, 1992), and SARSA (Rummery and Niranjan, 1994; Sutton, 1995), share a resemblance to the function of the striatal system (that guides habitual behavior). MB RL algorithms, such as DynaQ (Sutton and Barto, 1998), KWIK (Li et al., 2008), E3 (Kearns and Singh, 2002), R-Max (Brafman and Tennenholtz, 2002), and Learning with Opponent-Learning Awareness (Foerster et al., 2018), have a potential to arguably implement the function of the prefrontal cortex (that guides goal-directed behavior). In the human brain, the ventrolateral prefrontal cortex plays an important role in the meta-control among multiple systems. Inspired by this process, dynamic arbitration (Lee et al., 2014) algorithm has introduced the preliminary implementation of meta-control while meta-learning (Wang et al., 2016, 2018) emulated similar behavioral characteristics based on the MF RL approach. While using a context-aware, model-based RL to control the model-free to manage low-level skills is another viable solution of meta-control (Lee et al., 2009; Kulkarni et al., 2016; Hamrick et al., 2017), it doesn't appear to match with the prefrontal RL strategy (Lee et al., 2014; Wang et al., 2018; O'Doherty et al., 2021; Correa et al., 2022). A deeper investigation of the theory of RL in the human brain has not only inspired, but also justified the design of advanced RL algorithms (e.g., actor-critic, Barto et al., 1983) in addition to such progress. Rapid advances in deep neural network design enable the acceleration of such developments.

Such dramatic advances have resulted in the emergence of new unresolved issues. As discussed in Section 1, the fundamental principle of advanced algorithmic RLs is still (somewhat) far from that used in human RL. Figure 3B provides a good simple example highlighting this issue. While individual learning systems have been well-studied, the system(s) working at the meta-level (ventrolateral prefrontal cortex in this case), which is seen as a means to drive the optimal learning strategy subject to changes in goals and environments, has not been carefully taken into account.

We expect that an RL agent lacking the meta-control ability to integrate multiple learning strategies may not guarantee reliable prediction and adaptation performance. For example, an agent with an MB RL bias would start to make incorrect predictions when the measurement becomes noisy (i.e., due to high measurement noise or increasing uncertainty in the environmental structure) despite consuming a large amount of computing resources. Another possible scenario is that an RL agent with a fixed high learning rate (one-shot learning) may become unstable in learning about the environmental structure.

In practice, little investigation of the unified framework approach at the algorithmic level during RL has been performed. This is despite the fact that in the past few years, the neuroscientific community has made progress in understanding the neural basis of human intelligence. At the neuroscience level, progress has been made in identifying the neural circuit engaged during learning, as described in this paper. Converging evidence implicates the frontal pole as the core element of a second-order network able to read out the uncertainty associated with computations performed by other cortical circuits (Fleming et al., 2010; De Martino et al., 2013). Nevertheless, the lack of a clear algorithmic description of how meta-control appraisal interacts with learning has severely limited progress in our understanding of how striatal RL systems and the prefrontal meta-controller interact in an integrated single framework.

We firmly believe that the integration of new findings regarding MB/MF RL, and the use of rapid and slow RL in a single framework (as seen in Figure 3A) would be a natural resolution to the problems described above, as this is what our brains perform in daily life. First, supplementing MF RL with a functional module encapsulating MB RL would enable goal-directed decision-making based on a model of the environment. It will also enable rapid action selection to achieve

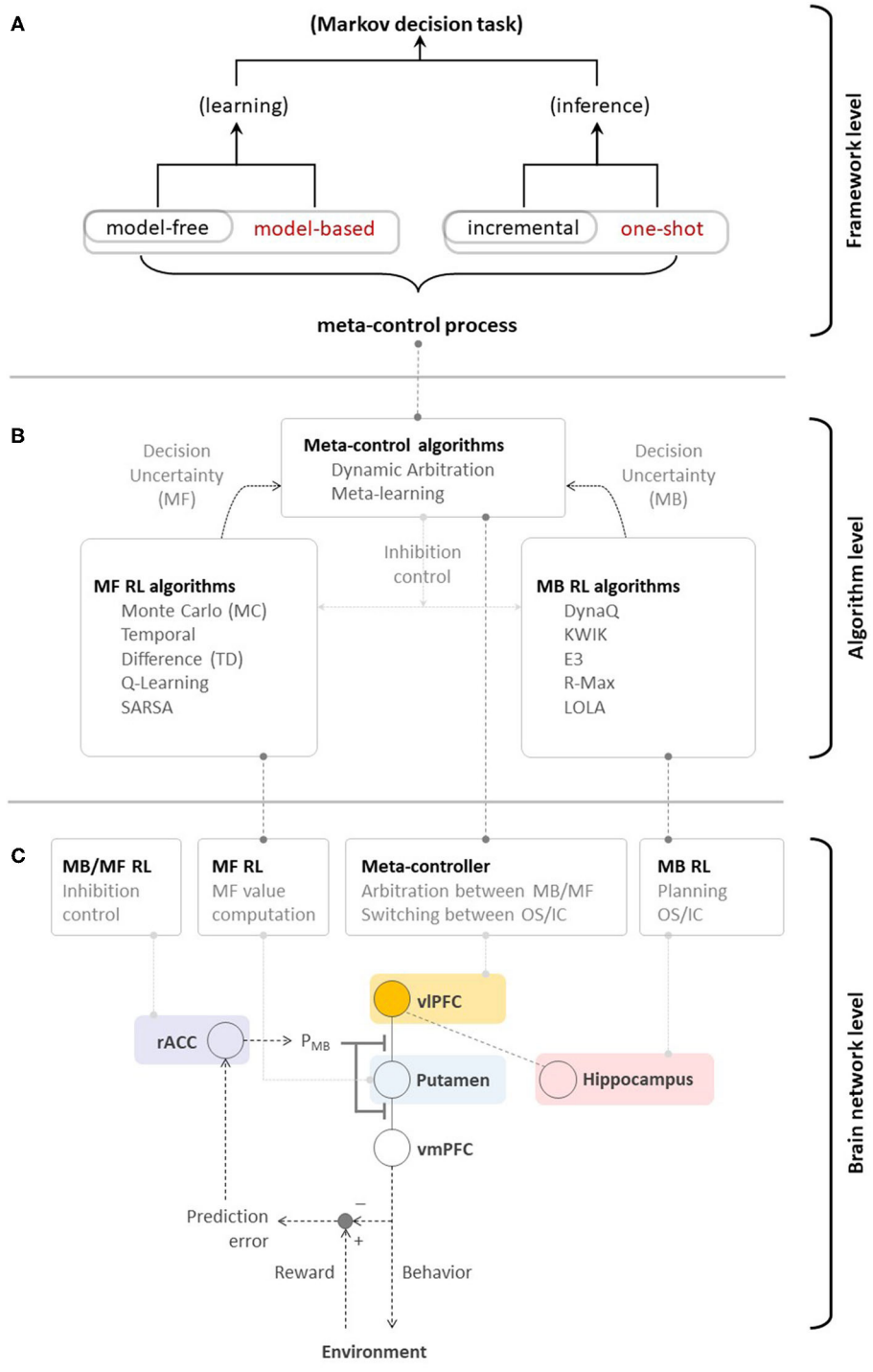

FIGURE 3
(A) Toward high performance, memory-efficient, and fast reinforcement learning—meta-control over MB/MF RL, and incremental/one-shot inference could ultimately accelerate the overall performance of algorithmic RL. **(B)** Neural evidence supporting algorithmic RL. **(C)** A brain network for human RL.

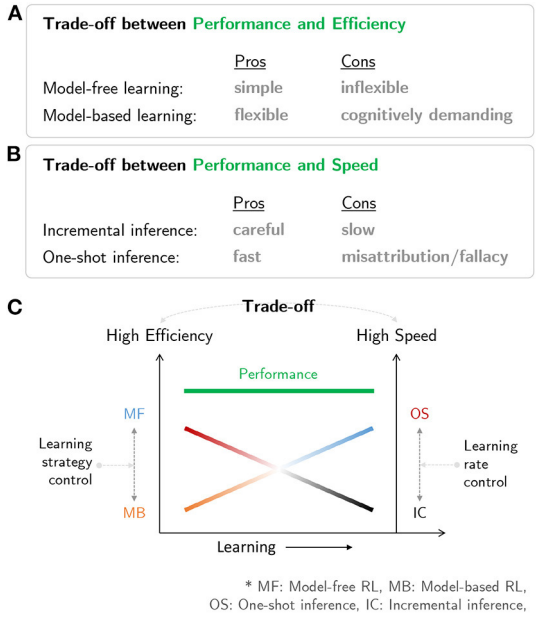

FIGURE 4
Characteristics of multiple systems in human RL: **(A)** Trade-off between performance and efficiency in MB/MF systems. **(B)** Trade-off between performance and speed in one-shot and incremental inference systems. **(C)** A neural mechanism in the human mind that resolves the trade-off issues among prediction performance, efficiency, and speed.

a goal in a dynamic environment. Second, supplementing an incremental learner with a one-shot learning module would ensure that RL agents would perform rapid learning based on information from a small number of episodes. Finally, implementing meta-control on these disparate learning strategies would afford us the leverage to rapidly achieve high prediction performance with a minimal loss of computational costs. Our view is supported by the recent evidence showing that in humans, the engagement of model-based RL mitigates the risk of assigning credit to outcome-irrelevant cues (Shahar et al., 2019). This result highlights the necessity of a control mechanism to determine when the MB system should override the MF system.

It is also noted that the meta-control has a great potential for dealing with conflicting demands in multiple agent learning, such as competition v.s. cooperation or envy v.s. guilt. As recent studies demonstrate that the MB and MF RL can provide pragmatic solutions for diverse social dilemma problems, implementing meta-control would also create a possibility for optimizing the performance of multi-agent learning systems. For example, such principles can be used to deal with a competition-cooperation issue in a smart home system in which multiple robot agents and human users interact with each other, or possibly to deal with an envy-guilt dilemma in social networks or online multiplayer games.

5. Conclusion

Deeply rooted in interdisciplinary research, including computer science, cognitive psychology, and decision neuroscience, reinforcement learning theories provide fundamental learning principles used to solve various real-world optimal control problems. In this paper, we reviewed a computational reinforcement learning theory, as well as its applications and challenges. We then discussed how the brain may perform analogous kinds of learning. We discussed MB/MF RL and one-shot/incremental inference, as well as meta-control over multiple learning strategies and discussed the implications of MB and MF RL in problems involving multiple interacting agents. We believe that the integrated conceptual framework incorporating these functions will lead to a major breakthrough in RL algorithm design where the trade-offs among prediction performance, computational costs, and time constraints are resolved.

Author contributions

JHL and SL contributed to the conception and design of the study. SA prepared the first draft of the figures. JHL wrote the first draft of the manuscript. JHL, JZL, and SL wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2019M3E5D2A01066267) and (No. 2020R1G1A1102683), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021M3E5D2A01022493), the Electronics and Telecommunications Research Institute(ETRI) grant funded by the Korean government (22ZS1100, Core Technology Research for Self-Improving Integrated Artificial Intelligence System), the Institute for Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (No. 2017-0-00451), and the Samsung Research Funding Center of Samsung Electronics under Project Number SRFCTC1603-52.

Conflict of interest

JZL is employed by DeepMind.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Axelrod, R., and Hamilton, W. D. (1981). The evolution of cooperation. *Science* 211, 1390–1396.

Badre, D., Doll, B. B., Long, N. M., and Frank, M. J. (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron* 73, 595–607. doi: 10.1016/j.neuron.2011.12.025

Balleine, B. W., and O'doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* 35, 48–69. doi: 10.1038/npp.2009.131

Barto, A. G., and Duff, M. (1994). "Monte Carlo matrix inversion and reinforcement learning," in *Advances in Neural Information Processing Systems*, 687–687.

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Trans. Syst. Man Cybernet.* 834–846.

Bellman, R. (1957). A Markovian decision process. *J. Math. Mech.* 679–684.

Bertsekas, D. P., and Tsitsiklis, J. N. (1995). "Neuro-dynamic programming: an overview." in *Proceedings of the 34th IEEE Conference on Decision and Control, 1995* (IEEE), 560–564.

Boorman, E. D., Behrens, T. E., Woolrich, M. W., and Rushworth, M. F. (2009). How green is the grass on the other side? frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron* 62, 733–743. doi: 10.1016/j.neuron.2009.05.014

Brafman, R. I., and Tennenholtz, M. (2002). R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.* 3, 213–231.

Busoniu, L., Babuska, R., De Schutter, B., and Ernst, D. (2010). *Reinforcement Learning and Dynamic Programming Using Function Approximators*, Vol. 39. CRC Press.

Camerer, C., Loewenstein, G., and Prelec, D. (2005). Neuroeconomics: how neuroscience can inform economics. *J. Econ. Lit.* 43, 9–64. doi: 10.1257/0022051053737843

Carroll, C., Cheng, P., and Lu, H. (2011). "Uncertainty and dependency in causal inference," in *Proceedings of the Cognitive Science Society*.

Christoph, G. R., Leonzio, R. J., and Wilcox, K. S. (1986). Stimulation of the lateral habenula inhibits dopamine-containing neurons in the substantia nigra and ventral tegmental area of the rat. *J. Neurosci.* 6, 613–619.

Claus, C., and Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI* 1998, 746–752.

Correa, C. G., Ho, M. K., Callaway, F., Daw, N. D., and Griffiths, T. L. (2022). Humans decompose tasks by trading off utility and computational cost. *arXiv preprint arXiv:2211.03890*.

Covington, P., Adams, J., and Sargin, E. (2016). "Deep neural networks for youtube recommendations," in *Proceedings of the 10th ACM Conference on Recommender Systems*, 191–198.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69, 1204–1215. doi: 10.1016/j.neuron.2011.02.027

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

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., and Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature* 441, 876–879. doi: 10.1038/nature04766

Dayan, P., and Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cogn. Affect. Behav. Neurosci.* 14, 473–492. doi: 10.3758/s13415-014-0277-8

Dayan, P., and Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cogn. Affect. Behav. Neurosci.* 8, 429–453. doi: 10.3758/CABN.8.4.429

De Jong, S., Tuyls, K., and Verbeeck, K. (2008). "Artificial agents learning human fairness," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems* (International Foundation for Autonomous Agents and Multiagent Systems), 863–870.

De Martino, B., Fleming, S. M., Garrett, N., and Dolan, R. J. (2013). Confidence in value-based choice. *Nat. Neurosci.* 16, 105–110. doi: 10.1038/nn.3279

Dickinson, A., Nicholas, D., and Adams, C. D. (1983). The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *Q. J. Exp. Psychol.* 35, 35–51.

Dolan, R. J., and Dayan, P. (2013). Goals and habits in the brain. *Neuron* 80, 312–325. doi: 10.1016/j.neuron.2013.09.007

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., and Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nat. Neurosci.* 18, 767–772. doi: 10.1038/nn.3981

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw.* 12, 961–974.

Doya, K., Samejima, K., Katagiri, K.-i., and Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural Comput.* 14, 1347–1369. doi: 10.1162/089976602753712972

Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., and Hester, T. (2021). Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Mach. Learn.* 110, 2419–2468. doi: 10.1007/s10994-021-05961-4

Evans, R., and Gao, J. (2016). *Deepmind AI Reduces Google Data Centre Cooling Bill by 40%*. DeepMind blog. Available online at: https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/

Fehr, E., and Camerer, C. F. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends Cogn. Sci.* 11, 419–427. doi: 10.1016/j.tics.2007.09.002

Fehr, E., and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Q. J. Econ.* 114, 817–868.

Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.* 28, 594–611. doi: 10.1109/TPAMI.2006.79

FitzGerald, T. H., Dolan, R. J., and Friston, K. J. (2014). Model averaging, optimal inference, and habit formation. *Front. Hum. Neurosci.* 8, 457. doi: 10.3389/fnhum.2014.00457

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., and Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science* 329, 1541–1543. doi: 10.1126/science.1191883

Floresco, S. B., West, A. R., Ash, B., Moore, H., and Grace, A. A. (2003). Afferent modulation of dopamine neuron firing differentially regulates tonic and phasic dopamine transmission. *Nat. Neurosci.* 6, 968–973. doi: 10.1038/nn1103

Foerster, J., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. (2018). "Learning with opponent-learning awareness," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (International Foundation for Autonomous Agents and Multiagent Systems), 122–130.

Fujimoto, S., Meger, D., and Precup, D. (2019). "Off-policy deep reinforcement learning without exploration," in *International Conference on Machine Learning* (PMLR), 2052–2062.

Garcia, V., and Bruna, J. (2018). "Few-shot learning with graph neural networks," in *International Conference on Learning Representations*.

Garety, P., Freeman, D., Jolley, S., Ross, K., Waller, H., and Dunn, G. (2011). Jumping to conclusions: the psychology of delusional reasoning. *Adv. Psychiatr. Treat.* 17, 332–339. doi: 10.1192/apt.bp.109.007104

Geisler, S., and Zahm, D. S. (2005). Afferents of the ventral tegmental area in the rat-anatomical substratum for integrative functions. *J. Comp. Neurol.* 490, 270–294. doi: 10.1002/cne.20668

Gesiarz, F., and Crockett, M. J. (2015). Goal-directed, habitual and Pavlovian prosocial behavior. *Front. Behav. Neurosci.* 9, 135. doi: 10.3389/fnbeh.2015.00135

Gintis, H. (2000). *Game Theory Evolving: A Problem-Centered Introduction to Modeling Strategic Behavior*. Princeton University Press.

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66, 585–595. doi: 10.1016/j.neuron.2010.04.016

Gläscher, J., Hampton, A. N., and O'doherty, J. P. (2008). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cereb. Cortex* 19, 483–495. doi: 10.1093/cercor/bhn098

Glimcher, P. W., and Fehr, E. (2013). *Neuroeconomics: Decision Making and the Brain*. Academic Press.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 3), 15647–15654. doi: 10.1073/pnas.1014269108

Griffiths, T. L., and Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychol. Rev.* 116, 661. doi: 10.1037/a0017201

Hampton, A. N., Bossaerts, P., and O'doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* 26, 8360–8367. doi: 10.1523/JNEUROSCI.1010-06.2006

Hamrick, J. B., Ballard, A. J., Pascanu, R., Vinyals, O., Heess, N., and Battaglia, P. W. (2017). "Metacontrol for adaptive imagination-based optimization," in *International Conference on Learning Representations*.

Hardin, G. (1968). The tragedy of the commons. *Science* 162, 1243–1248.

Hare, T. A., Camerer, C. F., and Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science* 324, 646–648. doi: 10.1126/science.1168450

Henrich, J. (2015). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton University Press.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., et al. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *Am. Econ. Rev.* 91, 73–78. doi: 10.1257/aer.91.2.73

Hessel, M., Danihelka, I., Viola, F., Guez, A., Schmitt, S., Sifre, L., et al. (2021). "Muesli: combining improvements in policy optimization," in *International Conference on Machine Learning* (PMLR), 4214–4226.

Holyoak, K. J., Lee, H. S., and Lu, H. (2010). Analogical and category-based inference: a theoretical integration with bayesian causal models. *J. Exp. Psychol. Gen.* 139, 702. doi: 10.1037/a0020488

Hughes, E., Leibo, J. Z., Philips, M. G., Tuyls, K., Duéñez-Guzmán, E. A., Castañeda, A. G., et al. (2018). Inequity aversion resolves intertemporal social dilemmas. *arXiv preprint arXiv:1803.08884*.

Jenkins, H. M., and Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychol. Monogr. Gen. Appl.* 79, 1.

Juechems, K., Balaguer, J., Ruz, M., and Summerfield, C. (2017). Ventromedial prefrontal cortex encodes a latent estimate of cumulative reward. *Neuron* 93, 705–714. doi: 10.1016/j.neuron.2016.12.038

Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., et al. (2018). "Scalable deep reinforcement learning for vision-based robotic manipulation," in *Conference on Robot Learning* (PMLR), 651–673.

Kearns, M., and Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Mach. Learn.* 49, 209–232. doi: 10.1023/A:1017984413808

Kim, S. H., and Lee, J. H. (2022). Evaluating SR-based reinforcement learning algorithm under the highly uncertain decision task. *KIPS Trans. Softw. Data Eng.* 11, 331–338.

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., and Tenenbaum, J. B. (2016). "Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction," in *COGSCI*.

Kollock, P. (1998). Social dilemmas: the anatomy of cooperation. *Annu. Rev. Sociol.* 24, 183–214.

Kulkarni, T. D., Narasimhan, K., Saeedi, A., and Tenenbaum, J. (2016). "Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation," in *Advances in Neural Information Processing Systems*, Vol. 29, eds D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc.).

Kuvayev, L., and Sutton, R. (1996). "Model-based reinforcement learning with an approximate, learned model," in *Proc. Yale Workshop Adapt. Learn. Syst*, 101–105.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behav. Brain Sci.* 40, e253. doi: 10.1017/S0140525X16001837

Lee, D., Seo, H., and Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annu. Rev. Neurosci.* 35, 287–308. doi: 10.1146/annurev-neuro-062111-150512

Lee, S. W., Kim, Y. S., and Bien, Z. (2009). A nonsupervised learning framework of human behavior patterns based on sequential actions. *IEEE Trans. Knowledge Data Eng.* 22, 479–492. doi: 10.1109/TKDE.2009.123

Lee, S. W., O'Doherty, J. P., and Shimojo, S. (2015). Neural computations mediating one-shot learning in the human brain. *PLoS Biol.* 13, e1002137. doi: 10.1371/journal.pbio.1002137

Lee, S. W., Shimojo, S., and O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron* 81, 687–699. doi: 10.1016/j.neuron.2013.11.028

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. (2017). "Multi-agent reinforcement learning in sequential social dilemmas," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems* (International Foundation for Autonomous Agents and Multiagent Systems), 464–473.

Lerer, A., and Peysakhovich, A. (2017). Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*.

Li, L., Littman, M. L., and Walsh, T. J. (2008). "Knows what it knows: a framework for self-aware learning," in *Proceedings of the 25th International Conference on Machine Learning* (ACM), 568–575.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., et al. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Littman, M. L. (1994). "Markov games as a framework for multi-agent reinforcement learning," in *Machine Learning Proceedings 1994* (Elsevier), 157–163.

Littman, M. L. (1996). *Algorithms for sequential decision making* (Ph.D. thesis). Brown University, Providence, RI, United States.

Matsumoto, M., and Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature* 447, 1111–1115. doi: 10.1038/nature05860

McAuliffe, K., Blake, P. R., Steinbeis, N., and Warneken, F. (2017). The developmental foundations of human fairness. *Nat. Hum. Behav.* 1, 42. doi: 10.1038/s41562-016-0042

McLaren, I., and Mackintosh, N. (2000). An elemental model of associative learning: I. latent inhibition and perceptual learning. *Anim. Learn. Behav.* 28, 211–246. doi: 10.3758/BF03200258

Meyniel, F., Schlunegger, D., and Dehaene, S. (2015). The sense of confidence during probabilistic learning: a normative account. *PLoS Comput. Biol.* 11, e1004305. doi: 10.1371/journal.pcbi.1004305

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., et al. (2016). "Asynchronous methods for deep reinforcement learning," in *International Conference on Machine Learning*, 1928–1937.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature* 518, 529–533. doi: 10.1038/nature14236

Moerland, T. M., Broekens, J., and Jonker, C. M. (2020). Model-based reinforcement learning: A survey. *arXiv preprint arXiv:2006.16712*.

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *J. Neurosci.* 16, 1936–1947.

Moore, S. C., and Sellen, J. L. (2006). Jumping to conclusions: a network model predicts schizophrenic patients' performance on a probabilistic reasoning task. *Cogn. Affect. Behav. Neurosci.* 6, 261–269. doi: 10.3758/CABN.6.4.261

Moutoussis, M., Bentall, R. P., El-Deredy, W., and Dayan, P. (2011). Bayesian modelling of jumping-to-conclusions bias in delusional patients. *Cogn. Neuropsychiatry* 16, 422–447. doi: 10.1080/13546805.2010.548678

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron* 41, 269–280. doi: 10.1016/S0896-6273(03)00869-9

Nasser, H. M., Chen, Y.-W., Fiscella, K., and Calu, D. J. (2015). Individual variability in behavioral flexibility predicts sign-tracking tendency. *Front. Behav. Neurosci.* 9, 289. doi: 10.3389/fnbeh.2015.00289

Niv, Y. (2009). Reinforcement learning in the brain. *J. Math. Psychol.* 53, 139–154. doi: 10.1016/j.jmp.2008.12.005

O'Doherty, J. P., Cockburn, J., and Pauli, W. M. (2017). Learning, reward, and decision making. *Annu. Rev. Psychol.* 68, 73–100. doi: 10.1146/annurev-psych-010416-044216

O'Doherty, J. P., Lee, S. W., and McNamee, D. (2015). The structure of reinforcement-learning mechanisms in the human brain. *Curr. Opin. Behav. Sci.* 1, 94–100. doi: 10.1016/j.cobeha.2014.10.004

O'Doherty, J. P., Lee, S. W., Tadayonnejad, R., Cockburn, J., Iigaya, K., and Charpentier, C. J. (2021). Why and how the brain weights contributions from a mixture of experts. *Neurosci. Biobehav. Rev.* 123, 14–23. doi: 10.1016/j.neubiorev.2020.10.022

OpenAI (2018). *OpenAI Five*. Available online at: https://blog.openai.com/openai-five/

Padoa-Schioppa, C., and Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature* 441, 223–226. doi: 10.1038/nature04676

Panait, L., and Luke, S. (2005). Cooperative multi-agent learning: the state of the art. *Auton. Agents Multiagent Syst.* 11, 387–434. doi: 10.1007/s10458-005-2631-2

Pearce, J. M., and Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* 87, 532.

Perolat, J., Leibo, J. Z., Zambaldi, V., Beattie, C., Tuyls, K., and Graepel, T. (2017). "A multi-agent reinforcement learning model of common-pool resource appropriation," in *Advances in Neural Information Processing Systems*, 3643–3652.

Pezzulo, G., Rigoli, F., and Chersi, F. (2013). The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Front. Psychol.* 4, 92. doi: 10.3389/fpsyg.2013.00092

Rangel, A., Camerer, C., and Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nat. Rev. Neurosci.* 9, 545–556. doi: 10.1038/nrn2357

Rapoport, A., Chammah, A. M., and Orwant, C. J. (1965). *Prisoner's Dilemma: A Study in Conflict and Cooperation*, Vol. 165. University of Michigan Press.

Recorla, R. A., and Wagner, A. R. (1972). "A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement," in *Classical Conditioning II: Current Research and Theory*, eds A. H. Black and W. F. Prokasy (New York, NY: Appleton- Century-Crofts), 64–99.

Roijers, D. M., Vamplew, P., Whiteson, S., and Dazeley, R. (2013). A survey of multi-objective sequential decision-making. *J. Artif. Intell. Res.* 48, 67–113. doi: 10.1613/jair.3987

Rummery, G. A., and Niranjan, M. (1994). *On-Line Q-Learning Using Connectionist Systems*. University of Cambridge, Department of Engineering.

Rushworth, M. F., Kolling, N., Sallet, J., and Mars, R. B. (2012). Valuation and decision-making in frontal cortex: one or many serial or parallel systems? *Curr. Opin. Neurobiol.* 22, 946–955. doi: 10.1016/j.conb.2012.04.011

Saez, A., Rigotti, M., Ostojic, S., Fusi, S., and Salzman, C. (2015). Abstract context representations in primate amygdala and prefrontal cortex. *Neuron* 87, 869–881. doi: 10.1016/j.neuron.2015.07.024

Schippers, M. C., and Van Lange, P. A. (2006). The psychological benefits of superstitious rituals in top sport: a study among top sportspersons 1. *J. Appl. Soc. Psychol.* 36, 2532–2553. doi: 10.1111/j.0021-9029.2006.00116.x

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., et al. (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature* 588, 604–609. doi: 10.1038/s41586-020-03051-4

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *J. Neurophysiol.* 80, 1–27.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599.

Shahar, N., Moran, R., Hauser, T. U., Kievit, R. A., McNamee, D., Moutoussis, M., et al. (2019). Credit assignment to state-independent task representations and its relationship with model-based decision making. *Proc. Natl. Acad. Sci. U.S.A.* 116, 15871–15876. doi: 10.1073/pnas.1821647116

Shenhav, A., Botvinick, M. M., and Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron* 79, 217–240. doi: 10.1016/j.neuron.2013.07.007

Si, J. (2004). *Handbook of Learning and Approximate Dynamic Programming*, Vol. 2. John Wiley & Sons.

Sigaud, O., and Buffet, O. (2013). *Markov Decision Processes in Artificial Intelligence*. John Wiley & Sons.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature* 529, 484–489. doi: 10.1038/nature16961

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* 362, 1140–1144. doi: 10.1126/science.aar6404

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014). "Deterministic policy gradient algorithms," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 387–395.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017b). Mastering the game of go without human knowledge. *Nature* 550, 354. doi: 10.1038/nature24270

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2017a). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.

Singh, S. P., and Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. *Mach. Learn.* 22, 123–158.

Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., et al. (2021). Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Mach. Learn.* 3, 9–44.

Sutton, R. S. (1995). "Generalization in reinforcement learning: Successful examples using sparse coarse coding," in *Advances in Neural Information Processing Systems*. p. 8.

Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*, Vol. 1. Cambridge: MIT Press.

Szepesvári, C. (2010). "Synthesis lectures on artificial intelligence and machine learning," in *Algorithms for Reinforcement Learning, Vol. 4* (Morgan and Claypool Publishers), 1–103.

Thibodeau, G. A., Patton, K. T., and Wills (1992). *Structure & Function of the Body*. St. Louis, MO: Mosby Year Book.

Thorndike, E. L. (1898). "Animal intelligence: An experimental study of the associative processes in animals," in *The Psychological Review: Monograph Supplements*. 2

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

Tricomi, E., Balleine, B. W., and O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *Eur. J. Neurosci.* 29, 2225–2232. doi: 10.1111/j.1460-9568.2009.06796.x

Tricomi, E., Rangel, A., Camerer, C. F., and O'Doherty, J. P. (2010). Neural evidence for inequality-averse social preferences. *Nature* 463, 1089. doi: 10.1038/nature08785

Valentin, V. V., Dickinson, A., and O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *J. Neurosci.* 27, 4019–4026. doi: 10.1523/JNEUROSCI.0564-07.2007

Van Hasselt, H., Guez, A., and Silver, D. (2016). "Deep reinforcement learning with double q-learning," in *AAAI*, 2094–2100.

van Otterlo, M., and Wiering, M. (2012). *Reinforcement Learning and Markov Decision Processes*. Berlin; Heidelberg: Springer.

Vecerik, M., Sushkov, O., Barker, D., Rothörl, T., Hester, T., and Scholz, J. (2019). "A practical approach to insertion with variable socket position using deep reinforcement learning," in *2019 International Conference on Robotics and Automation (ICRA)* (IEEE), 754–760.

Wan, Y., Rahimi-Kalahroudi, A., Rajendran, J., Momennejad, I., Chandar, S., and Van Seijen, H. H. (2022). "Towards evaluating adaptivity of model-based reinforcement learning methods," in *Proceedings of the 39th International Conference on Machine Learning*, eds K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato (PMLR), 22536–22561.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., et al. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* 21, 860. doi: 10.1038/s41593-018-0147-8

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., et al. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Watkins, C. J., and Dayan, P. (1992). Q-learning. *Mach. Learn.* 8, 279–292.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards* (Ph.D. thesis). University of Cambridge, Cambridge, United Kingdom.

Wunderlich, K., Dayan, P., and Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nat. Neurosci.* 15, 786–791.