Neural Networks 145 (2022) 80–89

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2021 Special Issue on AI and Brain Science: Perspective

Meta-learning, social cognition and consciousness in brains and machines

Angela Langdon[a], Matthew Botvinick[b,c], Hiroyuki Nakahara[d], Keiji Tanaka[d], Masayuki Matsumoto[e,f,g], Ryota Kanai[h,*]

[a] *Princeton Neuroscience Institute, Princeton University, USA*
[b] *DeepMind, London, UK*
[c] *Gatsby Computational Neuroscience Unit, University College London, London, UK*
[d] *RIKEN Center for Brain Science, Wako, Saitama, Japan*
[e] *Division of Biomedical Science, Faculty of Medicine, University of Tsukuba, Ibaraki, Japan*
[f] *Graduate School of Comprehensive Human Sciences, University of Tsukuba, Ibaraki, Japan*
[g] *Transborder Medical Research Center, University of Tsukuba, Ibaraki, Japan*
[h] *Araya, Inc. Tokyo, Japan*

ARTICLE INFO

Article history:
Available online 18 October 2021

Keywords:
Model-based reinforcement learning
Meta-learning
Social cognition
Consciousness

ABSTRACT

The intersection between neuroscience and artificial intelligence (AI) research has created synergistic effects in both fields. While neuroscientific discoveries have inspired the development of AI architectures, new ideas and algorithms from AI research have produced new ways to study brain mechanisms. A well-known example is the case of reinforcement learning (RL), which has stimulated neuroscience research on how animals learn to adjust their behavior to maximize reward. In this review article, we cover recent collaborative work between the two fields in the context of meta-learning and its extension to social cognition and consciousness. Meta-learning refers to the ability to learn how to learn, such as learning to adjust hyperparameters of existing learning algorithms and how to use existing models and knowledge to efficiently solve new tasks. This meta-learning capability is important for making existing AI systems more adaptive and flexible to efficiently solve new tasks. Since this is one of the areas where there is a gap between human performance and current AI systems, successful collaboration should produce new ideas and progress. Starting from the role of RL algorithms in driving neuroscience, we discuss recent developments in deep RL applied to modeling prefrontal cortex functions. Even from a broader perspective, we discuss the similarities and differences between social cognition and meta-learning, and finally conclude with speculations on the potential links between intelligence as endowed by model-based RL and consciousness. For future work we highlight data efficiency, autonomy and intrinsic motivation as key research areas for advancing both fields.

1. Introduction

Development of artificial intelligence (AI) and discoveries in neuroscience have been inspiration to each other. In this chapter, we discuss the current development of meta-learning in the intersection of AI and neuroscience. Meta-learning is particularly an interesting area of research to be discussed from both angles of AI and neuroscience, because it is considered as one of the key ingredients to build a more general form of AI that can learn to perform various tasks without having to learn them from scratch.

In this review, we will first introduce computational and empirical results in model-based reinforcement learning (RL) and illustrate their relevance to meta-learning both in AI and the brain (Section 2). We further illustrate new insights for meta-learning functions in the prefrontal cortex derived from deep learning implementation (Section 3). Furthermore, we point out commonalities and differences between meta-learning/meta-cognition and social cognition, emphasizing the importance of computational approaches to reveal them further (Section 4). Finally, we will conclude with speculations on the potential link between model-based meta RL and consciousness (Section 5).

A common thread across these topics is how both AI and the brain might benefit from the ability to utilize models to guide behavior and learning in the context of meta-learning, social cognition, and consciousness. As such, we aim to offer model-based reinforcement learning as a fundamental component of intelligence.

* Corresponding author.
E-mail address: kanair@araya.org (R. Kanai).

https://doi.org/10.1016/j.neunet.2021.10.004
0893-6080/© 2021 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

2. Model-based reinforcement learning for knowledge transfer

Reinforcement learning (RL) is a collection of algorithms designed to learn a behavioral policy (that is, rules for choosing actions) solely from ongoing experience of the rewarding or punishing consequences of those actions (Sutton & Barto, 1998). In RL, the impetus for learning is to maximize benefit or minimize cost and RL problems are typically formalized with respect to solving a single specific task: how to efficiently find a goal location in a spatial maze, or how to learn the best lever to press to trigger the later delivery of food. However, biological agents do not complete only one task in their lifetimes; humans and animals engage in many diverse tasks, across disparate timescales, both sequentially and in parallel. Learning to perform one task benefits from experience on related tasks, a process of *meta-learning*. Theories of meta-learning, whether from a biological, behavioral or artificial intelligence perspective, share the core idea that learning should exploit relevant past experience, rather than begin anew with each new task (Doya, 2002; Lemke et al., 2015; Vilalta & Drissi, 2002). Behaviorally, the signature of meta-learning is the speed-up of learning with repeated experience in a domain, this has been variously described as the formation of 'learning sets' (Harlow, 1949) or the integration of experience into structured knowledge known as a 'schema' (Bartlett, 1932; Tse et al., 2007).

The formal underpinning of RL relies on the concept of a *state*: policies and values are defined conditional on the current state, which is constructed from all relevant external and internal information to summarize the current configuration of a system (Minsky, 1967). For instance, the insertion of a lever into an operant chamber is a critical part of the current state of an instrumental task, and this event should be part of a state representation for efficient learning in this environment. But the notion of state is inherently flexible and critically depends on the current task and goals of an agent, who may exploit more or less knowledge about the generative statistics of the environment, their own current internal motivation and the outcomes of past learning experiences as required (Langdon et al., 2019; Nakahara & Hikosaka, 2012). This necessity to construct an appropriate state representation has led to the study of representation learning in biological and artificial systems: algorithms and architectures for forming a state representation in order to efficiently learn reward predictions and behavioral policies (Bengio, 2019; Nakahara, 2014; Niv, 2019).

Right from the outset, RL has benefitted from the flow of ideas between neuroscience and artificial intelligence, in its formalization and extension of behavioral theories of associative learning, which date in the psychological literature back to Pavlov and culminate in the influential Rescorla–Wagner model of trial-by-trial learning (Rescorla & Wagner, 1972). More recently, interest in RL theories of biological learning has exploded with the identification of phasic activity from midbrain dopamine (DA) neurons as a 'reward prediction error' (RPE) signal, the central teaching signal in RL, which putatively drives the update of value and/or policy representations in the brain (Eshel et al., 2015; Schultz et al., 1997; Watabe-Uchida et al., 2017). Careful study of the properties of DA prediction error signals across various tasks can illuminate the structure and content of the state representations that form the basis of reward predictions in behaving animals. One recent focus has centered on whether the reward predictions that drive DA RPE signals are essentially model-based or model-free (Akam & Walton, 2021; Dayan & Berridge, 2014; Langdon et al., 2018). In RL, the classic distinction between model-based and model-free algorithms speaks to the direct influence of a 'world model' in determining the likely transitions to future states, and thus the computation of future value. Model-free RL, on the other hand, relies on learning algorithms in which value estimates are cached, or stored, in the current state, rather than computed using an internal model of the task. Internal models may also sculpt learning through their influence on inferring the current 'hidden' state of the task in partially-observable environments (Rao, 2010; Yu, 2010). Many influential RL accounts of the activity of DA neurons during learning have been fundamentally model-free, typically assuming a temporal-difference (TD) learning rule (Sutton & Barto, 1990) over a representation that makes strict assumptions about the features, temporal characteristics and certainty of the progression of states encountered during a task (Ludvig et al., 2008; Montague et al., 1996; Schultz et al., 1997).

But recent evidence from DA RPE correlates has complicated these fundamental assumptions at the heart of TD models of reward learning in the brain, pointing to the role of internal task models in the formation and update of reward predictions in the brain (Nakahara et al., 2004). Foremost among these assumptions is that rewards can be substituted; TD learning algorithms are designed to aggregate reward amount irrespective of the 'packaging' in which that reward arrives. However, phasic DA responses are sensitive to unexpected changes in the flavor of a reward, suggesting it may also carry an identity- or state-prediction error signal alongside (or as a general version of) the more canonical RPE (Gardner et al., 2018; Takahashi et al., 2017). Second, TD learning requires experience of a state in order to update the cached value associated with that state. That is, TD learning has no mechanism by which value can transfer between predictive cues retrospectively. However, in a sensory preconditioning paradigm, DA RPE signals do indeed reflect such a retrospective transfer of value between cues, implying an internal model of the associations between cues is exploited for the formation of reward expectations in a new setting (Sadacca et al., 2016). Finally, classic TD learning models of the phasic DA response require a temporally precise state representation in order to achieve the temporal specificity observed in DA RPE correlates (Fiorillo et al., 2008; Hollerman & Schultz, 1998). Yet the assumption of time-point states required by these models precludes the generalization of value across moments in time. This is inconsistent with recent findings on the impact of VS lesions on the temporal specificity of DA RPE correlates, which suggest predictions about the timing and amount of upcoming rewards are neurally separable (Takahashi et al., 2016), and can be accounted for instead by an RL algorithm that learns about multiple properties of outcomes, including their timing and amount, in parallel.

In sum, these findings suggest that the reward learning circuitry in the brain exploits model-based learning algorithms in which multiple features of predicted outcomes are learned: their flavor, other associated cues and their timing (amongst others), in a way that allows them to be flexibly combined based on the current configuration of a task (Langdon et al., 2018). Learning in these algorithms is understood to be not just about 'value', but an array of features that provide separate sources of information about the broader structure of the 'reward landscape': how much attention is required to await the reward, in what form will it likely arrive, when should it be expected, how much effort needs to be expended in this environment in order to obtain it? While certainly more detailed than the simplest formulations of model-free TD learning, the internal model implied by model-based RL need not be prohibitively complex, sometimes requiring only a coarse approximation of the real environmental statistics (Akam et al., 2015; Park et al., 2020) (plus erie Boorman review to appear). Yet, even in simple conditioning tasks, which presumably require a relatively compact understanding of the contingent relationship between predictive cues and outcomes, phasic DA responses bear the striking hallmarks of inference in the computation of the current state, confirming an internal model of the task modulates reward predictions as they evolve during a trial (Starkweather et al., 2017, 2018).

What then are the implications of brain-inspired model-based RL for meta-learning in artificial learning systems? Model-based RL algorithms in which task representations are learned concurrently with reward expectations are naturally suited for meta-learning, in that any or all properties of the learned internal model of the tasks can be selectively generalized to a new setting, being accessible outside the reduced (model-free) construct of value (Fig. 1). An internal model that provides information about environmental regularities other than the strict reward-expectation estimator of model-free RL allows for the transfer (through inference) of this latent knowledge to a new setting. For example, building a model of the transition structure between states of the task, the different observation probabilities associated with these states, likely state durations and expectations about outcome identities can both facilitate learning in a single task environment (task A) and provide structured knowledge that can be applied to other tasks (here, task B and C). In this way, learning in a new task may be started with relevant priors about the contingent, temporal, effortful and attentional requirements of a whole class of problems, accelerating learning and supporting rapid behavioral adaptation to a novel environment.

3. Meta-learning in brains and machines

From the point of view of neuroscience, one of the most interesting recent developments in artificial intelligence is the rapid growth of deep reinforcement learning, the combination of deep neural networks with learning algorithms driven by reward (Botvinick et al., 2020). Since initial breakthrough applications to videogame tasks (Mnih et al., 2015), deep RL techniques have developed rapidly, with successful applications to much richer tasks (Schrittwieser et al., 2020; Vinyals et al., 2019), from complex motor control tasks (Merel et al., 2019) to multi-agent settings requiring both competition and cooperation (Jaderberg et al., 2019).

Deep RL brings together two computational frameworks with rich pre-existing ties to neuroscience research, but also shows how these can in principle be integrated, yielding artificial agents that – like biological agents – learn simultaneously how to represent the world and how to act adaptively within it (Botvinick et al., 2020). At the same time, however, dramatic differences have been noted between the way that standard deep RL agents learn, as compared with how humans and other animals learn. Most striking is an apparent difference in learning speed or 'sample efficiency': Deep RL often requires much more experience to arrive at an adaptive behavior than would a human learner (Botvinick et al., 2019; Lake et al., 2017).

While this difference in learning efficiency may appear to indicate a fundamental difference between biological learning and deep RL, it is worth considering the fact that human and animal learners gain their efficiency at least in part from the fact that they do not learn entirely 'from scratch.' Rather than operating as a tabula rasa, biological learners bring a wealth of past learning to bear on any new learning problem, and it is precisely their preexisting knowledge that enables rapid new learning. Psychologists have labeled this phenomenon learning to learn and meta-learning (Botvinick et al., 2019).

Recent research has begun to investigate whether meta-learning can be integrated into deep RL, with the same efficiency payoffs seen in biological learning. Studies of learning to learn in the machine learning context in fact date back at least to the 1990s (Thrun & Pratt, 1998), with methods explored both in deep learning (Hochreiter et al., 2001) and reinforcement learning research (Dayan, 1993). However, it is only recently that the topic has been studied in the context of deep RL. Over the past three years, though, meta-learning in deep RL – or as it is increasingly called 'meta-reinforcement learning' – has become a burgeoning area of AI research.

Many studies of meta-reinforcement learning propose novel learning algorithms. Among these, a general motif is 'taking gradients of gradients', that is, adjusting hyper-parameters of a network (for example, parameters governing its initial state or the operation of its weight-update procedure) by making changes in a direction that would reduce some measure of error, such as the temporal-difference error in a reinforcement learning algorithm (Finn et al., 2017; Xu et al., 2018).

While such approaches are potentially powerful, a different and more minimalistic approach may, we believe, have more immediate relevance to neuroscience. Here, no special, dedicated mechanism is installed to support meta-learning. Instead, meta-learning emerges spontaneously from more basic mechanisms. The two ingredients that are necessary are (1) a learning system that has some form of short-term memory, and (2) a training environment that exposes the learning system not to a single task, but instead to a sequence or distribution of interrelated tasks. When these two ingredients are simultaneously present, something remarkable occurs: The system slowly learns to use its short-term memory as a basis for fast learning (Santoro et al., 2016; Wang et al., 2016).

To make this concrete, consider a recurrent neural network, trained using reinforcement learning on a series of 'bandit' tasks. Recurrent networks are, of course, endowed with a form of short-term memory, since their recurrent connectivity supports maintenance of information over time through sustained patterns of unit activity. The precise information that is so maintained, and the way such information is represented, are of course determined by the network's connection weights. In deep RL, those weights are slowly adjusted through (typically) gradient-descent learning. In the bandit scenario, such learning leads to weight adjustments which, in turn, allow the network to 'keep around' in its activation patterns information about past actions and rewards, precisely the information needed to adapt appropriately to each new bandit problem. In sum, slow (weight-based) learning gives rise to a much faster (activity-based) learning algorithm (Botvinick et al., 2019; Duan et al., 2016; Santoro et al., 2016; Wang et al., 2016).

This emergent form of meta-reinforcement learning bears some intriguing parallels with neuroscience. In particular, the two forms of memory that make it work, weight-based and activation based, map to synapse-based and activation based or 'working' memory in the brain. Indeed, the notions that both synapse-based and working memory subserve reinforcement learning, and that synapse-based learning serves to regulate the function of working memory, have been deeply explored in computational neuroscience (Chatham & Badre, 2015; Collins & Frank, 2012). These themes bear a particularly strong connection to the function of prefrontal cortex, given its putative role in working memory (Nakahara & Hikosaka, 2012).

With these connections in mind, Wang and colleagues (Wang et al., 2018) proposed a theory of prefrontal cortex (PFC) function, based on emergent meta-reinforcement learning. The theory anchors on two functional–anatomical aspects of the PFC: (1) its strong recurrent connectivity, supporting working memory through sustained neural activation, (2) its participation in a circuit running through the striatum, where dopamine plays a pivotal role in modulating synaptic strength based on reward. Computationally, these aspects of PFC correspond to the recurrent connectivity and RL-based weight adjustment that drive emergent meta-reinforcement learning, as studied in AI research. With this in mind, Wang and colleagues (Wang et al., 2018) presented

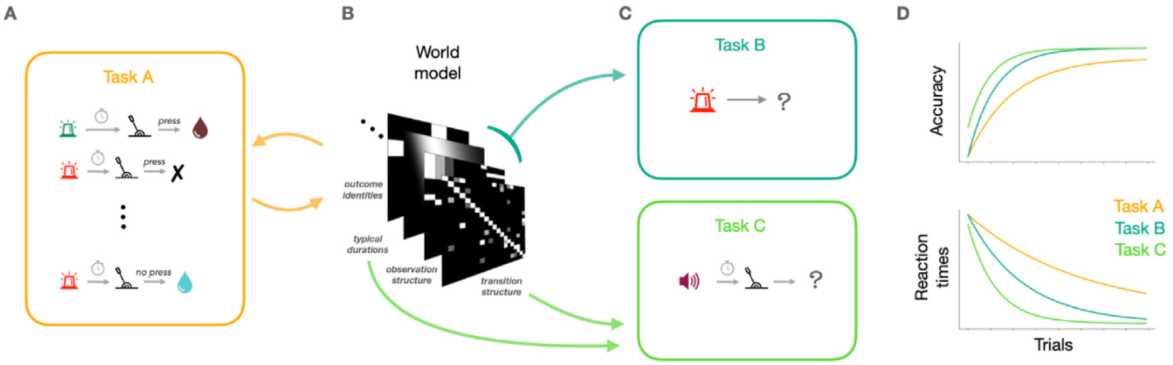

Fig. 1. Model-based reinforcement learning as a mechanism for meta-learning. (A) Model-based reinforcement learning algorithms allow for the acquisition and application of structured knowledge about the environment during experience on a task. (B) This knowledge comprises a world model, that summarizes the learned probabilistic relationship between successive states of the environment and how they depend on actions (transition structure), observations and cues in the environment and the underlying hidden state (observation structure), temporal characteristics of task states (for e.g., their typical duration) and the relationship between states and specific form of either rewarding or aversive outcomes (outcome identities). Other structured knowledge, such as required effort, attentional demands and so on, may also be learned. (C) This model-based knowledge can act as a prior, sculpting expectations and actions even in entirely novel environments. Either the entire world model, or selective parts of this knowledge, may be applied to accelerate learning in new tasks. (D) The behavioral signatures of meta-learning are variously faster acquisition of correct responses, increased asymptotic accuracy or higher initial accuracy, along with faster initial and increasingly rapid responding with experience on collections of tasks.

simulations demonstrating how emergent meta-RL can parsimoniously explain a range of otherwise poorly understood experimental observations from the PFC and dopamine literatures. These ranged from behavioral patterns reported in the earliest work on learning-to-learn to Bayes-optimal adjustments in learning rate mediated by PFC representations, to single-unit coding properties in bandit-like problems, to patterns of dopamine activity reflecting sensitivity to task structure. A particularly striking result showed that patterns of behavior and dopamine activity that had previously been attributed to 'model-based' reinforcement learning mechanisms could in fact arise from 'model-free' reinforcement learning, in a meta-RL setting. This observation, which has since been echoed in the AI literature (Guez et al., 2019), opens up a new space of computational possibilities between pure model-free and model-based learning, providing a new perspective on the varieties of RL operative in biological systems (Botvinick et al., 2019).

In subsequent work building on the proposal from Wang and colleagues (Wang et al., 2018), Ritter and colleagues studied how emergent meta-RL might interact with an additional episodic memory mechanism (Botvinick et al., 2019). Drawing on neuroscience data concerning contextual reinstatement effects in prefrontal cortex, Ritter and colleagues (Ritter et al., 2018) showed how meta-RL might sculpt episodic memory function in a way that supports rapid retrieval of previously discovered solutions, when previously solved problems recur after an intervening delay. This extension of the meta-RL theory bears rich connections with recent work indicating a central role for episodic memory in human reinforcement learning (Gershman & Daw, 2017), opening up new avenues for investigating this territory.

One important limitation of meta-RL research, in both the AI and neuroscience realms, has been the simplicity of the tasks to which it has been applied. A key question is whether the sorts of mechanisms involved in emergent meta-RL (or indeed other algorithmic varieties of meta-RL and meta-learning in general) are sufficient to support the discovery of abstract structural principles, akin to those that humans work with in naturalistic domains ranging from cooking to driving to computer use to scientific research (Lake et al., 2017). In recent work, Wang and colleagues (Wang et al., 2021) have introduced a video-game benchmark, nicknamed "Alchemy", which is specifically designed to test the conceptual reach of current methods for meta-RL. Initial results, reported by these investigators, suggest that current techniques do hit an identifiable 'glass ceiling', in terms of their ability to discover and exploit abstract task structure. If this limitation can be confirmed through convergent research, then it suggests that new innovations in meta-learning may be required to push this upper bound. If so, then human abilities for structure learning, and their neural underpinnings, may serve as a useful guide for further advances on the AI front.

4. Commonalities between social cognition and meta-learning

In this section, we discuss the relationship between social cognition/learning and meta-cognition/learning. Meta-cognition refers to additional, self-referential processes that work on top of primary cognition and provides a foundation for meta-learning. For instance, assessing the degree of confidence in making one's own decision and adapting the learning rate in response to environmental volatility are considered meta-cognition and meta-learning, respectively (Doya, 2002; Fleming et al., 2012; Yeung & Summerfield, 2012). Social cognition and social learning are summarized in a similar manner as additional or self-referential processes that modulate and improve cognition and learning in social contexts. For instance, decision making in a social context requires taking into accounts not only gains of the self but also of others (Fehr, 2009; Rilling & Sanfey, 2011). Also, while learning is typically associated with one's own experience, it can be improved additionally learning from observing others' experience of their choices and outcomes (Burke et al., 2010; Cooper et al., 2012; Suzuki et al., 2012).

While meta-learning and social learning are discussed in different contexts, they both play contribute to learning processes. Social contexts presuppose considerations of others such as predicting their behavior, and assessments of the effects of one's own actions on others and predicting what is in the mind of other individuals. These socio-cognitive functions are considered as the learning and inference of latent variables of other individuals. That is, while the mental states of others are not directly observable from their behavior, we regard them as underlying the observable behavior and as such we can model them as latent variables. The learning and inference of those latent variables adds to learning in social contexts. Whether social cognition and learning involve self-referential processes as in meta-learning has been a research question in this research area.

This question has been addressed by a study by Suzuki et al. (2012). They investigated how human participants learn to predict the value-based decisions by simulating other people's decision-making processes (Suzuki et al., 2012). Using fMRI combined with computational modeling, they found that two types of learning signals are essential for this task. One is called simulated-others' reward prediction error (sRPE), which is the difference between the outcomes to others and the expected reward to others, generated by simulation of value-based decision-making for others. The other signal is called simulated-others' action prediction error (sAPE), which is the discrepancy between the choice of others and the expected choice (probability) of others, generated by the simulated decision-making process. The sRPE keeps track of others assuming a common decision-making process between the self and others, while the sAPE corrects the deviations from the expectation for others' actual behavior, reflecting the differences in the decision-making processes. Furthermore, they found a form of self-referential property for the internal simulation both for the learning signals and for the decision signals. As the simulation theory in social cognition suggests, brain areas associated with the sRPE (i.e., ventromedial prefrontal cortex) also encoded signals for one's own decisions and for predicting decisions of others. These results answer the question above by supporting the notion that social cognition involves self-referential processes (see Fig. 2).

Humans and other social animals often exhibit behavior for the benefit of other individuals. For humans, it has been debated whether such prosocial behavior originates from (ultimately) self-regarding or from truly others' regarding. Regardless of the theoretical position, cognition of others' decision making seems to involve conversion of others' benefit or value into the self-oriented decision-making process. Fukuda et al. (2019) investigated neural underpinnings of this social value conversion (Fukuda et al., 2019). They assessed how a bonus reward for others is embedded in one's own value-based decisions, compared to the case where the bonus was offered to the participants themselves. They found that the bonus offer was processed for both others and the self in left dorsolateral prefrontal cortex (ldlPFC), but uniquely for others in the right temporoparietal junction (rTPJ). The influence of the bonus for others on one's own decision was processed in the right anterior insula (rAI) and in the vmPFC. Using dynamic causal modeling and psycho-physiological interaction analysis, they showed influences from the ldlPFC and rTPJ to the rAI and from the rAI to the vmPFC, suggesting a neural cascade of the social value conversion, rTPJ/ldlPFC → rAI → vmPFC. Furthermore, this neural mechanism of social value conversion was found to be different between selfish and prosocial individuals. This conversion process is not meta-cognition per se but can still be regarded as similar in its function to module and improve self-regarding decision-making in social settings.

As we have seen in the two studies above, research in social cognition and learning asks how social information (i.e., social cues from others) modulates social functions. For example, our interpretation of other people's actions differ depending on our knowledge of their intention (Cooper et al., 2010). On the other hand, metacognition and meta-learning are studied in a more general context. An arising question is whether social functions are realized by dedicated social brain circuits or supported by more general brain circuits. From evolutionary perspectives, one might argue that the brain has adapted to social situations and developed dedicated social brain circuits. Alternatively, one could also argue that social functions may be embedded within more general brain circuits. Whether and what social functions may be viewed as part of general cognitive functions remains to be studied in future research.

In this section, we discussed the relationship between social cognition and meta-cognition. In social cognition, there are cases in which social functions lead to undesirable consequences. For example conformity may be beneficial for harmonious relationships in a group. However, it could lead to suboptimal choices. Similarly, if we step back from meta and social functions and look broadly at cognition and learning, there are abundant examples suggesting that our behavior may be suboptimal. Such examples include risk aversion and regret aversion, and a wide range of cognitive heuristics (Gilovich et al., 2002; Tversky & Kahneman, 1974). It may be worth investigating meta and social functions not only in relation to their benefits but also to their limitations or sub-optimality, since such findings would reveal underlying mechanisms.

Finally, in analogy with the historical role of computational studies in propelling research into neural mechanisms underlying RL, we expect that computational studies should shed a new light on social cognition and meta-cognition/learning. As discussed in earlier sections, RL frameworks have begun with studies of conditioning (Sutton & Barto, 1990, 1998), followed by the finding of dopamine neural responses associated with reward prediction error (Montague et al., 1996; Schultz et al., 1997), leading now to a much wider domains of cognition, decision-making and learning (Behrens et al., 2009; Dayan, 2012, 2012; Dayan & Nakahara, 2018; Kim & Hikosaka, 2015; Montague et al., 2012, 2006). Fertile fields are open for computational studies that investigate the benefits and associated limitations of meta-cognition and social cognition from computational perspectives.

5. Consciousness and intelligence

Consciousness and intelligence have been treated as distinct notions, because there is no a priori reason to believe that one depends on the other. In this section, we challenge this view and discuss possible links between consciousness and intelligence. Specifically, we discuss the hypothesis that consciousness evolved as a platform for meta-learning underlying general intelligence. In this section, we discuss the functional aspects of consciousness (i.e., access consciousness) rather than phenomenal experience (or qualia), namely, observable cognitive functions that humans and animals perform when they are conscious as opposed to when they cannot report the content of consciousness.

Here, we provisionally use the term "general intelligence" as the ability to efficiently solve multiple tasks, including tasks novel to the agent, using knowledge and models learned through past experiences. With this definition, intelligence is measured by the efficiency of meta-learning and transfer learning. This definition is in line with the various previous notions of intelligence such as the formal definition by Hutter and colleagues (Hutter, 2000; Legg & Hutter, 2007; Leike & Hutter, 2018) and more recently by Chollet (2019).

Here, we discuss the following two possible methods that the brain might use to solve a broader range of tasks than those directly trained on and argue that they are linked with possible functions of consciousness. The first method is internal simulation using pre-trained forward models of the world. This is in line with the model-based approach to meta-learning in RL (see Fig. 1). The second method is to combine previously learned models in a flexible manner to build a solution to novel tasks. While these are by no means the only ways to build general intelligence, we focus on these solutions to novel tasks and discuss how those functions are related to consciousness.

The first approach is to use internal simulation to find a policy to a new task using world models learned through interactions with the environment. Even when we are presented with new goals, the models of the dynamics of the environment remains the same, and an agent can use the models for internal simulation to figure out how to solve a task in their imagination (Ha & Schmidhuber, 2018; Hafner et al., 2020).

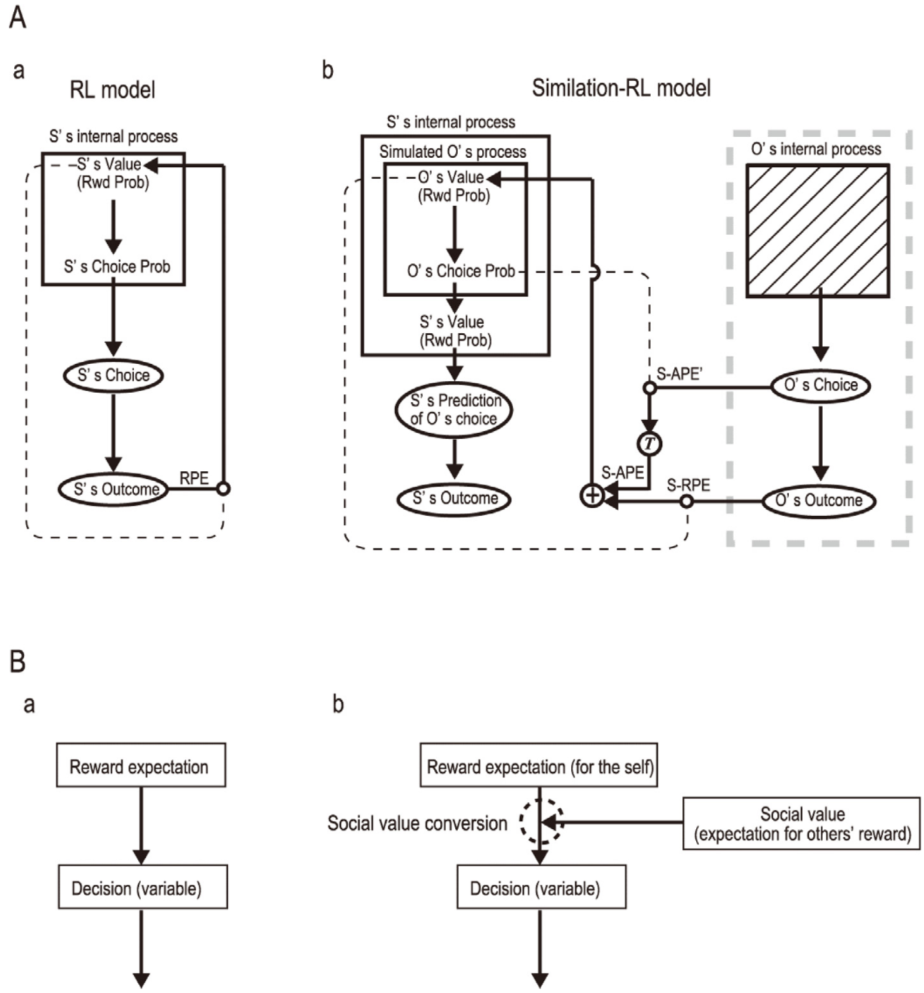

Fig. 2. Schematic diagram for social learning and decision-making process. (A) Reinforcement learning (RL) model (a) and Simulation learning rooted on internal process of reinforcement learning (b), adopted from supplemental Figure 1 (Suzuki et al., 2012). (a) Box indicates the subjects' (S's) internal decision making process. When the outcome is presented, the value of the chosen option (or the stimulus reward probability) is updated, using reward prediction error (RPE: discrepancy between S's value and actual outcome). (b) Decision making process of subjects during the Other task (Suzuki et al., 2012) is modeled by Simulation-RLsRPE+sAPE (S-RLsRPE+sAPE) model. The large box on the left indicates the subject's internal process; the smaller box inside indicates the other's (O's) internal decision making process being simulated by the subject. The large box on the right, outlined by a thick dashed line, corresponds to what the other is 'facing in this task', and is equivalent to what subjects were facing in the Control task (compare with the schematic in (a)). The hatched box inside corresponds to the other's internal process, which is hidden from the subjects. As modeled by the S-RLsRPE+sAPE, at the time of decision, subjects use the learned simulated-other's value to first generate the simulated-other's choice probability (O's Choice Prob), based on which they generate their own value (S's Value) and the subject's choice probability for predicting the other's choice (S's Choice Prob). Accordingly, subjects then predict the other's choice. Once the outcome is shown, subjects update the simulated-other's value using the simulated-other's reward and action prediction errors (sRPE and sAPE), respectively; sRPE is the discrepancy between the simulated-other's value and the other's actual outcome, and sAPE is the discrepancy between the simulated-other's choice probability and the other's actual choice, in the value level. The simulated-other's action prediction error is first generated in the action level (denoted by sAPE' in the figure) and transformed (indicated by *T* in the open circle) to the value level, becoming the sAPE to update the simulated-other's value, together with the sRPE. (B). (a) Decision-making is driven by value (reward expectation) when not involving other individuals. (b) By contrast, in social situations, value (reward expectation for the self) is complemented by social value (reward expectation for other individuals); for this complement, social value needs a conversion to be merged with original value (called social value conversion). They together drive decision-making.

While it remains a matter of debate what functions consciousness adds to biological systems, there are a few experimental situations known to require consciousness for successful performance. A typical example is an experimental condition called trace conditioning in classical conditioning (Clark et al., 2002; Clark & Squire, 1998; Droege et al., 2021; Knight et al., 2006). In the trace conditioning of the eye-blink, there is a non-overlapping temporal gap between a conditioned stimulus (CS) such as a tone and a following unconditioned stimulus (US) such as an air puff to the eyes. Contrary to the trace conditioning, in delay conditioning, there is a temporal overlap between the offset of CS and the onset of US. If the subject successfully keeps expecting US, they close the eyelid in a non-reflective manner. Empirical evidence suggests that only trace conditioning require consciousness for successful responses to CSs whereas delay conditioning did not (Clark et al., 2002; Clark & Squire, 1998; Droege et al., 2021; Knight et al., 2006). These findings suggest that retention of information over time involves consciousness.

Another clue comes from experiments on the agnosia patient DF who had impairments in conscious object recognition (Goodale et al., 1991). When she was asked to indicate the orientation of a slanted slit verbally or by adjusting a handle, she could not report the orientation, suggesting that she had no awareness of the orientation. However, she could post a letter through the slit by adjusting the orientation of the letter in the right angle, suggesting that she could use the orientation information for guiding action. This is a classic example that led to the proposal that the ventral pathway (where DF had a damage) is needed for conscious vision, whereas the dorsal pathway guides action without evoking conscious experience. Crucially, when she was shown the slit first, and then the light was turned off so that she would have to wait for a few seconds before acting, then she failed to reach the slit correctly (Goodale et al., 1994).

These experimental examples suggest that the unconscious action system needs to be guided online by auditory or visual information and to act on offline information retained from the recent past, which requires the conscious perception of the cue tone or the shape. In other words, "online systems" that process information real time works without consciousness, but to maintain information over time, consciousness is necessary. This information maintenance can be one functional benefit of consciousness.

By analyzing what is common among cognitive tasks that seem to require awareness, Kanai et al. (2019) proposed the Information Generation Theory (IGT) of Consciousness. This theory proposes that a function of consciousness is to internally generate sensory representations of the past or the future event, not happening in the present, thus allowing to bridge the temporal gap.

According to the view presented there, interactions with internal generated representations allow an agent to perform a variety of non-reflexive behaviors associated with consciousness such as cognitive functions enabled by consciousness such as intention, imagination, planning, short-term memory, attention, curiosity, and creativity. Furthermore, the hypothesis suggests that consciousness emerged in evolution when organisms gained the ability to perform internal simulations using generative models. This characterization of consciousness is in essence in line with the functions of model-based RL and links a potential function of consciousness to a possible mechanism of general intelligence implemented in biological systems.

The second approach to general intelligence is to combine pretrained models in a flexible manner to establish a solution to a new problem. Neural networks are functions that convert input vectors into output vectors and as such can be combined in a flexible manner as long as they have corresponding dimensionality. Even if a new problem requires a transformation that cannot be handled by a single function, a combination of previously learned functions could produce the required transformation. For example, consider a neural network f: $x \rightarrow y$ that outputs class classification y from an image x and another network g: $y \rightarrow z$ that converts the label of the class y to an audio output z. Both f and g are neural networks that solve specific problems, but we can compose $g \circ f$ to establish a new solution that converts an input image x to a speech signals z. When many pre-trained networks are available for such flexible combinations, one can configure a vast number of new networks simply by combining them (Fig. 3). From the viewpoint of accomplishing general intelligence, new tasks can be solved by finding a path from one node to another in a directed graph consisting of pre-trained neural networks.

To allow flexible combinations of neural networks, those networks need to be linked together via a shared representation. Otherwise, those networks are disjoint and cannot be combined. To date, functions of consciousness have not been discussed much in terms of data compatibility across modalities in the brain. In a recent paper, VanRullen and Kanai re-interpreted the Global Workspace Theory (GWT) in terms of shared representations across multiple, specialized modules in the brain (VanRullen & Kanai, 2021). Briefly, GWT is a cognitive model of consciousness (Baars, 1997, 2005; Dehaene et al., 1998) in which the brain is divided into specialized modules for specific functions and long-distance connections that connect those specialized modules. In GWT, the outputs from the fast, specialized modules are broadcast and shared among those distinct modules. The network of modules where information is shared is called the global workspace and is thought to allow the system to solve a new problem by coordinating information coming from specialized modules. This process allows flexible operations on the shared information via slow and effortful processes. These fast and slow processes roughly correspond to system 1 and system 2 modes of thinking (Kahneman, 2013).

This is a re-interpretation and extension of the original GWT offers a possible link between consciousness and intelligence. As argued above, this architecture allows construction of a vast number of new functions by combining existing neural networks in a flexible manner. One can conceive consciousness as a platform that enables flexible combinations of specialized modules (Bengio, 2019). While this insight does not tell us whether an AI system with global workspace has subjective experience, it provides an inspiration for building AI systems that can adapt to new tasks by applying existing knowledge from past experiences.

Moreover, this interpretation of GWT makes an example where consideration of implementation of cognitive models with modern deep learning architectures helps refine cognitive concepts. For example, the precise operation of broadcasting in GWT has been open to multiple interpretations when one tries to computationally implement it. Here, more concrete views emerge once we consider the global workspace as shared latent space. As such, it was argued that considering computational implementations of cognitive models with deep learning is useful for clarifying sometimes vague concepts and processes discussed in cognitive theories. Similar approaches should be applicable for other known cognitive or psychological models.

Previously, Dehaene et al. (2017) proposed that two types of information processing are associated with consciousness. One is the selection of information for broadcasting across the system for flexible applications for various purposes. The other is metacognition to estimate the uncertainty of first-order processing such as perception and memory (Shea & Heyes, 2010). In this section, while we focused on the broadcasting function, we did not explicitly discuss the role of metacognition in consciousness. However, it is an important direction to consider the role of metacognition in this context, as there has been a proposal that links global workspace to metacognition (Shea & Frich, 2019).

In this section, we discussed possible links between consciousness and intelligence and considered model-based approaches as a common principle underlying both consciousness and intelligence. This insight came from consideration of possible implementations of known cognitive functions with modern deep learning architectures. This approach would be a fruitful direction for the intersection of neuroscience and machine learning.

6. Future directions

In this review, we discussed how model-based RL in machine learning informs neuroscience and vice versa in the context of meta-learning, social cognition and consciousness. While biologically inspired approaches to AI have been often discussed, recent work indicates that our notions of cognitive modules are also refined by applications of AI developments for interpreting brain functions.

There are a few research areas that would benefit from the interdisciplinary approach. One is the issue of data efficiency in AI research. While deep neural networks have been successful in the ability to recognize and classify images, they tend to require huge datasets of annotated (i.e., labeled) images to learn discriminative features. The human visual system has the capability to learn from much fewer examples. Lake et al. argued that such human-like learning machines should (1) build causal models of the world that support explanation and understanding, rather than merely solving pattern recognition problems; (2) ground learning in intuitive theories of physics and psychology to support and

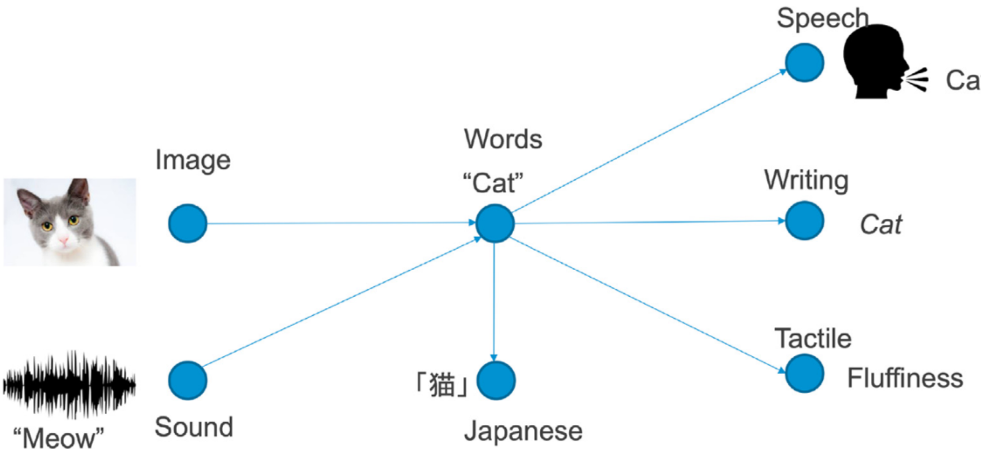

Fig. 3. An illustration of solving a broader range of tasks by combining pre-trained neural networks. Each arrow represents a neural network trained on a specific task. The figure illustrates the idea that global workspace enables flexible combination of pretrained models. For example, the arrow from the image of a cat to the word 'cat' represents an image recognition neural network, and another arrow from the world 'cat' to the text image of Cat (on the right) represents a network transforming text data into text images. Once we have many pretrained networks, solving a new task corresponds to finding a path in the network of pretrained networks. A more elaborated treatment of this idea has been discussed in VanRullen and Kanai (2021).

enrich the knowledge that is learned; and (3) harness compositionality and learning-to-learn to rapidly acquire and generalize knowledge to new tasks and situations (Lake et al., 2017). These are all important features for meta-learning needed to accomplish successful learning from small sample.

Yet another domain where meta-learning in neuroscience informs AI research is implementation of autonomy. Even in modern AI systems, goals and reward functions are often hand-crafted by human engineers, but this is crucial for AI systems to continuously learn from the environment. Higher mammals including humans are endowed with the ability voluntarily generate future goals and reward functions and continue to learn throughout their lifetime. One possible research direction for creating autonomy is the intrinsic motivations such as curiosity and empowerment (Klyubin et al., 2005; Oudeyer, 2007). People engage in actions that do not directly lead to survival or reproduction and rather lead to "their own sake" (Baldassarre et al., 2014; Berlyne, 1966). Various algorithms have been already proposed in AI research and they are meta-cognitive in the sense that curiosity is driven by expected prediction error. Data efficiency, autonomy and intrinsic motivation are all linked with meta-cognition/learning and do have practical implications for extending the horizon of current AI technologies. These shared themes between neuroscience and AI research will be key for driving discoveries in both fields.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Japan Society for the Promotion of Science (KAKENHI Grants 16H06570 and 16K21738 to H.N., 16H06564 to K.T., and 16H06567 to M.M.).

References

Akam, T., Costa, R., & Dayan, P. (2015). Simple plans or sophisticated habits? State, transition and learning interactions in the two-step task. *PLoS Computational Biology*, *11*(12), Article e1004648. http://dx.doi.org/10.1371/journal.pcbi.1004648.

Akam, T., & Walton, M. E. (2021). What is dopamine doing in model-based reinforcement learning? *Current Opinion in Behavioral Sciences*, *38*, 74–82. http://dx.doi.org/10.1016/j.cobeha.2020.10.010.

Baars, B. J. (1997). In the theatre of consciousness: Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, *4*, 292–309.

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in Brain Research*, *150*, 45–53. http://dx.doi.org/10.1016/S0079-6123(05)50004-9.

Baldassarre, G., Stafford, T., Mirolli, M., Redgrave, P., Ryan, R. M., & Barto, A. (2014). Intrinsic motivations and open-ended development in animals, humans, and robots: an overview. *Frontiers in Psychology*, *5*, 985. http://dx.doi.org/10.3389/fpsyg.2014.00985.

Bartlett, F. C. (1932). *Remembering; a Study in Experimental and Social Psychology*. Cambridge [Eng.]: The University Press.

Behrens, T. E., Hunt, L. T., & Rushworth, M. F. (2009). The computation of social behavior. *Science*, *324*, 1160–1164. http://dx.doi.org/10.1126/science.1169694.

Bengio, Y. (2019). The consciousness prior. arXiv preprint arXiv:1709.08568 [cs.LG].

Berlyne, D. E. (1966). Curiosity and exploration. *Science*, *153*, 25–33. http://dx.doi.org/10.1126/science.153.3731.25.

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*, 408–422. http://dx.doi.org/10.1016/j.tics.2019.02.006.

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. *Neuron*, *107*, 603–616. http://dx.doi.org/10.1016/j.neuron.2020.06.014.

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 14431–14436. http://dx.doi.org/10.1073/pnas.1003111107.

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

Chollet, F. (2019). On the measure of intelligence. arXiv preprint arXiv:1911.01547 [cs.AI].

Clark, R. E., Manns, J. R., & Squire, L. R. (2002). Classical conditioning, awareness, and brain systems. *Trends in Cognitive Sciences*, *6*(12), 524–531. http://dx.doi.org/10.1016/S1364-6613(02)02041-7.

Clark, R. E., & Squire, L. R. (1998). Classical conditioning and brain systems: The role of awareness. *Science*, *280*(5360), 77–81. http://dx.doi.org/10.1126/science.280.5360.77.

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

Cooper, J. C., Dunne, S., Furey, T., & O'Doherty, J. P. (2012). Human dorsal striatum encodes prediction errors during observational learning of instrumental actions. *Journal of Cognitive Neuroscience*, *24*, 106–118. http://dx.doi.org/10.1162/jocn_a_00114.

Cooper, J. C., Kreps, T. A., Wiebe, T., Pirkl, T., & Knutson, B. (2010). When giving is good: Ventromedial prefrontal cortex activation for others' intentions. *Neuron*, *67*, 511–521. http://dx.doi.org/10.1016/j.neuron.2010.06.030.

Dayan, P. (1993). Improving generalization for temporal difference learning - the successor representation. *Neural Computation*, *5*, 613–624. http://dx.doi.org/10.1162/neco.1993.5.4.613.

Dayan, P. (2012). Twenty-five lessons from computational neuromodulation. *Neuron*, *76*, 240–256. http://dx.doi.org/10.1016/j.neuron.2012.09.027.

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(2), 473–492. http://dx.doi.org/10.3758/s13415-014-0277-8.

Dayan, P., & Nakahara, H. (2018). Models and methods for reinforcement learning. In Wagenmakers E.-J. (Ed.), *Methodology: vol. 5, Stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 507–546). John Wiley & Sons, Inc..

Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 14529–14534.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 486–492.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*(4), 495–506. http://dx.doi.org/10.1016/S0893-6080(02)00044-8.

Droege, P., Weiss, D., Schwob, N., & Braithwaite, V. (2021). Trace conditioning as a test for animal consciousness: a new approach. *Animal Cognition*, *24*, http://dx.doi.org/10.1007/s10071-021-01522-3.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016). RL2: Fast reinforcement learning via slow reinforcement learning. arXiv preprint arXiv:1611.02779 [cs.AI].

Eshel, N., Bukwich, M., Rao, V., Hemmelder, V., Tian, J., & Uchida, N. (2015). Arithmetic and local circuitry underlying dopamine prediction errors. *Nature*, *525*(7568), 243–246. http://dx.doi.org/10.1038/nature14855.

Fehr, E. (2009). Social preferences and the Brain. In P. W. Glimcher, C. Camerer, R. A. Poldrack, & E. Fehr (Eds.), *Neuroeconomics decision making and the Brain* (pp. 215–232). Academic Press.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. arXiv preprint arXiv:1703.03400 [cs.LG].

Fiorillo, C. D., Newsome, W. T., & Schultz, W. (2008). The temporal precision of reward prediction in dopamine neurons. *Nature Neuroscience*, *11*(8), 966–973. http://dx.doi.org/10.1038/nn.2159.

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *Journal of Neuroscience*, *32*, 6117–6125. http://dx.doi.org/10.1523/JNEUROSCI.6489-11.2012.

Fukuda, H., Ma, N., Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., Cheng, K., & Nakahara, H. (2019). Computing social value conversion in the human brain. *Journal of Neuroscience*, *39*, 5153–5172. http://dx.doi.org/10.1523/JNEUROSCI.3117-18.2019.

Gardner, M. P. H., Schoenbaum, G., & Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B*, *285*(1891), Article 20181645, https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2018.1645.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, *68*, 101–128. http://dx.doi.org/10.1146/annurev-psych-122414-033625.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: the psychology of intuitive judgment*. Cambridge University Press.

Goodale, M. A., Jakobson, L. S., & Keillor, J. M. (1994). Differences in the visual control of pantomimed and natural grasping movements. *Neuropsychologia*, *32*(10), 1159–1178. http://dx.doi.org/10.1016/0028-3932(94)90100-7.

Goodale, M. A., Milner, A. D., Jakobson, L. S., & Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature*, *349*(6305), 154–156. http://dx.doi.org/10.1038/349154a0.

Guez, A., Mirza, M., Gregor, K., Kabra, R., Racanière, S., Weber, T., Raposo, D., Santoro, A., Orseau, L., Eccles, T., Wayne, G., Silver, D., & Lillicrap, T. (2019). An investigation of model-free planning. arXiv preprint arXiv:1901.03559 [cs.LG].

Ha, D., & Schmidhuber, J. (2018). World models. http://dx.doi.org/10.5281/zenodo.1207631, arXiv preprint arXiv:1803.10122 [Cs, Stat].

Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. (2020). Dream to control: Learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603 [Cs]. http://arxiv.org/abs/1912.01603.

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, *56*(1), 51–65. http://dx.doi.org/10.1037/h0062474, APA PsycArticles.

Hochreiter, S., Younger, A. S., & Conwell, P. R. (2001). Learning to learn using gradient descent. In *International conference on artificial neural networks* (pp. 87–94). Springer.

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *1*(4), 304–309. http://dx.doi.org/10.1038/1124.

Hutter, M. (2000). A theory of universal artificial intelligence based on algorithmic complexity. arXiv preprint arXiv:cs/0004001 [cs.AI].

Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castaneda, A. G., Beattie, C., Rabinowitz, N. C., Morcos, A. S., Ruderman, A., Sonnerat, N., Green, T., Deason, L., Leibo, J. Z., Silver, D., Hassabis, D., Kavukcuoglu, K., & Graepel, T. (2019). Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, *364*, 859–865. http://dx.doi.org/10.1126/science.aau6249.

Kahneman, D. (2013). *Thinking, fast and slow* (1st ed.). Farrar, Straus and Giroux.

Kanai, R., Chang, A., Yu, Y., Magrans de Abril, I., Biehl, M., & Guttenberg, N. (2019). Information generation as a functional basis of consciousness. *Neuroscience of Consciousness*, *2019*(1), http://dx.doi.org/10.1093/nc/niz016, niz016.

Kim, H. F., & Hikosaka, O. (2015). Parallel basal ganglia circuits for voluntary and automatic behaviour to reach rewards. *Brain*, *138*, 1776–1800. http://dx.doi.org/10.1093/brain/awv134.

Klyubin, E. S., Polani, D., & Nehaniv, C. L. (2005). Empowerment: A universal agent-centric measure of control. In *IEEE Congress on Evolutionary Computation*. IEEE.

Knight, D. C., Nguyen, H. T., & Bandettini, P. A. (2006). The role of awareness in delay and trace fear conditioning in humans. *Cognitive, Affective, & Behavioral Neuroscience*, *6*(2), 157–162. http://dx.doi.org/10.3758/CABN.6.2.157.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, Article e253. http://dx.doi.org/10.1017/S0140525X16001837.

Langdon, A. J., Sharpe, M. J., Schoenbaum, G., & Niv, Y. (2018). Model-based predictions for dopamine. *Current Opinion in Neurobiology*, *49*, 1–7. http://dx.doi.org/10.1016/j.conb.2017.10.006.

Langdon, A. J., Song, M., & Niv, Y. (2019). Uncovering the 'state': Tracing the hidden state representations that structure learning and decision-making. *Behavioural Processes*, *167*, Article 103891. http://dx.doi.org/10.1016/j.beproc.2019.103891.

Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. In *Advances in artificial general intelligence: concepts, architectures and algorithms: proceedings of the AGI workshop 2006* (pp. 17–24).

Leike, J., & Hutter, M. (2018). On the computability of solomonoff induction and AIXI. *Theoretical Computer Science*, *716*, 28–49. http://dx.doi.org/10.1016/j.tcs.2017.11.020.

Lemke, C., Budka, M., & Gabrys, B. (2015). Metalearning: A survey of trends and technologies. *Artificial Intelligence Review*, *44*(1), 117–130. http://dx.doi.org/10.1007/s10462-013-9406-y.

Ludvig, E. A., Sutton, R. S., & Kehoe, E. J. (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Computation*, *20*(12), 3034–3054. http://dx.doi.org/10.1162/neco.2008.11-07-654.

Merel, J., Botvinick, M., & Wayne, G. (2019). Hierarchical motor control in mammals and machines. *Nature Communications*, *10*, 5489. http://dx.doi.org/10.1038/s41467-019-13239-6.

Minsky, M. L. (1967). *Computation: finite and infinite machines*. Prentice-Hall, Inc.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529–533. http://dx.doi.org/10.1038/nature14236.

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

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, *16*, 72–80. http://dx.doi.org/10.1016/j.tics.2011.11.018.

Montague, P. R., King-Casas, B., & Cohen, J. D. (2006). Imaging valuation models in human choice. *Annual Review of Neuroscience*, *29*, 417–448. http://dx.doi.org/10.1146/annurev.neuro.29.051605.112903.

Nakahara, H. (2014). Multiplexing signals in reinforcement learning with internal models and dopamine. *Current Opinion in Neurobiology*, *25*, 123–129. http://dx.doi.org/10.1016/j.conb.2014.01.001.

Nakahara, H., & Hikosaka, O. (2012). Learning to represent reward structure: a key to adapting to complex environments. *Neuroscience Research*, *74*, 177–183. http://dx.doi.org/10.1016/j.neures.2012.09.007.

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

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

Oudeyer, P.-Y. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, *1*(6), http://dx.doi.org/10.3389/neuro.12.006.2007.

Park, S. A., Miller, D. S., Nili, H., Ranganath, C., & Boorman, E. D. (2020). Map making: Constructing, combining, and inferring on abstract cognitive maps. *Neuron*, *107*(6), 1226–1238. http://dx.doi.org/10.1016/j.neuron.2020.06.030, e8.

Rao, R. P. N. (2010). Decision making under uncertainty: A neural model based on partially observable Markov decision processes. *Frontiers in Computational Neuroscience*, *4*, http://dx.doi.org/10.3389/fncom.2010.00146.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II: current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Rilling, J. K., & Sanfey, A. G. (2011). The neuroscience of social decision-making. *Annual Review of Psychology*, *62*, 23–48. http://dx.doi.org/10.1146/annurev.psych.121208.131647.

Ritter, S., Wang, S., Kurth-Nelson, Z., Jayakumar, S., Blundell, C., Pascanu, R., & Botvinick, M. (2018). Been there, done that: meta-learning with episodic recall. In *International conference on machine learning*.

Sadacca, B. F., Jones, J. L., & Schoenbaum, G. (2016). Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *ELife*, *5*, http://dx.doi.org/10.7554/eLife.13665.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. In *International conference on machine learning* (pp. 1842–1850).

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., & Silver, D. (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, *588*, 604–609. http://dx.doi.org/10.1038/s41586-020-03051-4.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. http://dx.doi.org/10.1126/science.275.5306.1593.

Shea, N., & Frich, C. D. (2019). The global workspace needs metacognition. *Trends in Cognitive Sciences*, *23*(7), 560–571.

Shea, N., & Heyes, C. (2010). Metamemory as evidence of animal consciousness: the type that does the trick. *Biology and Philosophy*, *25*(1), 95–110.

Starkweather, C. K., Babayan, B. M., Uchida, N., & Gershman, S. J. (2017). Dopamine reward prediction errors reflect hidden-state inference across time. *Nature Neuroscience*, *20*(4), 581–589. http://dx.doi.org/10.1038/nn.4520.

Starkweather, C. K., Gershman, S. J., & Uchida, N. (2018). The medial prefrontal cortex shapes dopamine reward prediction errors under state uncertainty. *Neuron*, *98*(3), 616–629. http://dx.doi.org/10.1016/j.neuron.2018.03.036, e6.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel, & J. Moore (Eds.), *Learning and computational neuroscience: foundations of adaptive networks* (pp. 497–537). Cambridge, MA.: MIT Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction* (1st ed.). MIT Press.

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., Cheng, K., & Nakahara, H. (2012). Learning to simulate others' decisions. *Neuron*, *74*, 1125–1137. http://dx.doi.org/10.1016/j.neuron.2012.04.030.

Takahashi, Y. K., Batchelor, H. M., Liu, B., Khanna, A., Morales, M., & Schoenbaum, G. (2017). Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. *Neuron*, *95*(6), 1395–1405. http://dx.doi.org/10.1016/j.neuron.2017.08.025.

Takahashi, Y. K., Langdon, A. J., Niv, Y., & Schoenbaum, G. (2016). Temporal specificity of reward prediction errors signaled by putative dopamine neurons in rat VTA depends on ventral striatum. *Neuron*, *91*(1), 182–193. http://dx.doi.org/10.1016/j.neuron.2016.05.015.

Thrun, S., & Pratt, L. (1998). *Learning to learn*. Springer Science & Business Media.

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P., & Morris, R. G. M. (2007). Schemas and memory consolidation. *Science*, *316*(5821), 76. http://dx.doi.org/10.1126/science.1135935.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131. http://dx.doi.org/10.1126/science.185.4157.1124.

VanRullen, R., & Kanai, R. (2021). Deep learning and the global workspace theory. *Trends in Neuroscience*, *44*(9), 692–704.

Vilalta, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial Intelligence Review*, *18*(2), 77–95. http://dx.doi.org/10.1023/A:1019956318069.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., ... Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, *575*, 350–354. http://dx.doi.org/10.1038/s41586-019-1724-z.

Wang, J. X., King, M., Porcel, N., Kurth-Nelson, Z., Zhu, T., Deck, C., Choy, P., Cassin, M., Reynolds, M., Song, F., Buttimore, G., Reichert, D. P., Rabinowitz, N., Matthey, L., Hassabis, D., Lerchner, A., & Botvinick, M. (2021). Alchemy: A structured task distribution for meta-reinforcement learning. arXiv preprint arXiv:2102.02926 [Cs]. http://arxiv.org/abs/2102.02926.

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*, 860–868. http://dx.doi.org/10.1038/s41593-018-0147-8.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., & Botvinick, M. (2016). Learning to reinforcement learn. arXiv preprint arXiv:1611.05763 [cs.LG].

Watabe-Uchida, M., Eshel, N., & Uchida, N. (2017). Neural circuitry of reward prediction error. *Annual Review of Neuroscience*, *40*(1), 373–394. http://dx.doi.org/10.1146/annurev-neuro-072116-031109.

Xu, Z. W., van Hasselt, H., & Silver, D. (2018). Meta-Gradient Reinforcement Learning. In *Advances in neural information processing systems (vol. 31)*.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *367*, 1310–1321. http://dx.doi.org/10.1098/rstb.2011.0416.

Yu, S.-Z. (2010). Hidden semi-Markov models. *Artificial Intelligence*, *174*(2), 215–243. http://dx.doi.org/10.1016/j.artint.2009.11.011.