HHS Public Access

Author manuscript

Nat Rev Neurosci. Author manuscript; available in PMC 2021 January 11.

Published in final edited form as:

Nat Rev Neurosci. 2020 October ; 21(10): 576–586. doi:10.1038/s41583-020-0355-6.

Beyond simple dichotomies in reinforcement learning.

Anne GE Collins[1,*], **Jeffrey Cockburn**[2]

[1]University of California, Berkeley, Department of Psychology, Berkeley, CA 94712, USA

[2]California Institute of Technology, Division of the Humanities and Social Sciences, Pasadena, CA 91125, USA

Abstract

Reinforcement learning (RL) is a framework of particular importance to psychology, neuroscience, and machine learning. Interactions between these fields, as promoted through the common hub of RL, has facilitated paradigm shifts relating multiple levels of analysis within a singular framework (e.g dopamine function). Recently, more sophisticated RL algorithms have been incorporated to better account for human learning, and in particular its oft documented reliance on two separable systems. However, along with many benefits, this dichotomous lens can distort questions, and may contribute to an unnecessarily narrow perspective on learning and decision making. Here we outline some of the consequences that come from over-confidently mapping algorithms, such as model-based vs. model-free RL, with putative cognitive processes. We argue that the field is well positioned to move beyond simplistic dichotomies, and we propose a means of re-focusing research questions toward the rich and complex components that comprise learning and decision making.

1 Introduction

The empirical study of learning and decision making, in both human and non-human animals, has catalogued a wealth of evidence consistent with the idea that behavior is governed by at least two separable controllers. Behavior has been dichotomized across several dimensions, including emotion (Hot/Cold)[1], action selection (habitual/goal-directed)[2], judgements (associative vs. rule based)[3], and more recently, model-free/model-based (MF/MB)[4]. Although the terms used to characterize these controllers vary, and have largely been absorbed into the terms System1/System2[5,6], many seemingly 'irrational' behaviours have been argued to emerge from a system that is fast, reactive, implicit, retrospective and emotionally charged. This has been contrasted with a system described as slow, deliberative, explicit, prospective and calculating[5,6]. Our understanding of the processes driving behavior, from the neural implementations to social factors, has advanced considerably through the use of these dichotomies in terms of both experimental and theoretical development.

* annecollins@berkeley.edu.

Competing interests

The authors declare no competing interests

However, despite a common philosophical core, the various frameworks used to describe these behavioral controllers vary in terms of their formalism and scope, and as such, they are not interchangeable, nor are the phenomena they purport to explain. More importantly, the aforementioned dichotomies do not constrain the neural or cognitive mechanisms that dissociate the two systems, making it deceptively difficult to uniquely and reliably classify behavior as being driven by any one particular controller. To address this, dual-system theories of learning and decision making have been drawn toward the formalization offered by the field of machine learning, readily found in the literature as a mapping to model-based (MB) / model-free (MF) reinforcement learning (RL)[7].

Computational formalization promises important benefits: it promotes a precise quantitative definition of important concepts, and often enable us to bridge levels of analysis[8] across cognitive concepts to their underlying neural mechanisms. Parameters of formal computational models are often thought to capture meaningful information about how we learn, in a low-dimensional and easily quantifiable (parameter) space. While the MB/MF RL formalization has realized such benefits[9], it has also brought some challenges[10]. Here we address some of the limitations presented by dual-system theories that have the potential to impede progress in the associated fields of study. We argue that the dimensionality of learning – the axes of variance that describe how individuals learn and make choices – is well beyond two, as proposed by any given dual-system theory. We contend that attempts to better understand learning and decision making by mapping it onto two a-priori defined components may cause the field to lose sight of some essential features of learning. We focus on the example of the MB vs. MF RL dichotomy for one key reason: MB vs. MF is one of the most well-defined dichotomous theories of learning and decision-making, and has often been interpreted as capturing the essence of other dual-system theories computationally. We show that this confidence induced by a strong formalism does not obviate the limitations of the dual-system approach. Although the strengths offered by the MB/MF RL framework are well documented[9,11], it has become increasingly clear that accurately labelling behavior or neurological signals as uniquely associated with one algorithm or the other can be deceptively difficult[12–16]. Here, we address some of the MB/MF framework's limitations, highlighting sources of misattribution, the challenges associated with aligning computational and mechanistic primitives, and what is lost when our theoretical lens is narrowed to a single dimension. We propose that refocusing on the computationally defined primitives of learning and decision making that bridge brain and behavior may offer a more fruitful path forward.

2 What is Reinforcement learning?

Reinforcement learning (RL) is a term widely used in at least three separate, though overlapping, fields of research: computational sciences (machine-learning, artificial intelligence, computer science); behavioral sciences (psychology, cognitive sciences); and neuroscience (systems, cellular) (fig. 1). Although use of a shared language has mutually enriched these three disciplines, slight conceptual distinctions can lead to confusion across the three domains. In computational settings, RL refers to a class of learning environments and algorithms in which learning is driven by a scalar value (the *reinforcement*) and the algorithm's goal is to optimize the future cumulative reinforcement (see box 1 for details).

Behavioral sciences use RL in reference to learning processes that promote behavior by pairing it with a valued outcome (or the removal of an undesired outcome), and discouraged otherwise. The field of neuroscience typically treats RL as a process through which neuronal pathways are shaped in the brain, and is most often thought of as dopamine-dependent plasticity that shapes learning within and between various brain regions (including cortico-striatal networks).

2.1 RL algorithms

Computational RL defines a class of learning problems and algorithms such as model-free (MF) and model-based (MB) RL. In contrast to supervised learning where correct answers are provided, or unsupervised learning where no feedback is available at all, RL problems involve learning how to achieve a goal by using the rewards and punishments induced through interactions with the environment. The family of RL algorithms is defined by their objective function: to find a strategy that maximizes the accumulated future reward. Some tasks are simple enough that an RL approach can solve the learning problem completely by identifying the best actions from start to finish in all possible scenarios (e.g. playing tic-tac-toe). However, most real world problems (like driving to work) are far more complex: the number of possible circumstances in which the agent might find itself (the state space) can be huge, as are the number of available actions, while measures of progress can be murky. In cases such as this, RL algorithms are limited to learning how to make ‘good’ decisions as opposed to completely solving what is often an intractable problem.

A formal description of an RL problem consists of a set of states in which the learning agent might find itself, and a set of actions the agent can take. It also includes a transition function that describes how the environment will respond to the agent’s actions, and a reward function that defines how good (or bad) observed events are. It is important to note that a formal specification is, as with any model, an approximation of the real problem. Most RL algorithms decompose decision making into two steps: first, derive value estimates for the different states or action available, then choose actions that are deemed most valuable.

RL algorithms can be categorized along many dimensions. MB vs. MF algorithms are contrasted based on the extent to which they represent the environment. MB algorithms maintain a representation of the problem beyond the state and action space, usually the transition and reward function. Equipped with a task model, the agent guides its decisions by considering the consequences of its actions to construct a plan that will move it toward its goal. Model-free (MF) algorithms, as their name implies, do not maintain such an explicit model. Instead, they store a set of value estimates, each representing the aggregated reward history of choices selected by the agent in the past, from which the algorithm can gauge the expected benefit of the options on offer (see Box 1).

These two strategies can be contrasted with respect to how they respond to changes in the environment or the agent’s goal. MB algorithms adapt more readily as they can leverage their task model to dynamically plan toward an arbitrary goal, though they suffer the additional hindrance of computing this action plan, which can rapidly become intractable. MF algorithms cannot adapt as easily due to their strategy of integrating reward history into a single value estimate; however, they offer an efficient approach to learning and decision

making. Consider, for example, a student arriving at the main cafeteria for lunch, where they unexpectedly find a stand offering samples from a new cafe on campus (fig. 2). In contrast to the bland offerings from the cafeteria, the sample food is fresh and delicious and would clearly be a better lunch option. The next day the student considers their meal options. A MB strategy would consult its map of the campus and the items available to devise a plan of action that would route the student to the new cafe for lunch. In contrast, a MF strategy would consult its value estimates and simply repeat yesterday's choice to visit the cafeteria since that option has been rewarding in the past, particularly after the last visit. In contrast to the potentially complex, and often intractable planning problem faced by a MB agent, MF choice is considerably less effortful as it relies on a cached value estimate that can be derived using simple computation that rely only on easily accessible information (see box 2) signalling how "off" the current estimate is. However, the computational efficiency of a MF approach causes it to be relatively inflexible as it can only look to the past to inform its choices, whereas the prospective capacity of the MB agent[17] allows it to flexibly adapt to changes in the environment or its own goals.

The scientific progress resulting from applying a RL computational framework is plainly apparent through the rapid advances in cognitive neuroscience. RL has been pivotal in providing a sound quantitative theory of learning, and a normative framework through which we can understand the brain and behavior. As an explanatory framework, RL advances our understanding beyond phenomenology in ascribing functional structure to observed data. Here we highlight some of the key findings.

2.2 MF-RL and the brain

Early research into the principles that govern learning likened behavior to the output of a stimulus-response association machine that builds links between stimuli and motor responses through reinforcement[18]. Various models described the relationship between stimuli, response, and reward, with nearly all sharing a common theme of an associative process driven by a surprise signal[19–21]. Computational reinforcement learning theory built on the principles animal behaviorists had distilled through experimentation to develop the method of temporal difference (TD) learning (a model-free algorithm), which offers general purpose learning rules while also formalizing the reinforcement learning problem[22].

The TD-RL algorithm sparked a turning point in our understanding of dopamine function in the brain. In a seminal set of studies, the phasic firing patterns of dopamine (DA) neurons in the ventral tegmental area (VTA) were shown to mirror the characteristics of a TD-RL reward prediction error (see equ 1), offering a bridge between behaviorally descriptive models and a functional understanding of a learning algorithms embodied by the brain[23–25]. Continued work along this line of research has probed the details of DA activity in greater detail, linking it to various flavors of MF-RL[26,27]. Importantly, this work has shifted the conceptualization of S-R instrumental learning away from inflexible reflex-like behaviour toward one of adaptable value based learning.

The role of DA as a MF-RL teaching signal is supported by work in both human and non-human animals showing that DA affects cortico-striatal plasticity as predicted by the theory[28]. Subsequent research has focused on the causal import of dopaminergic input to

show that systematic modulation of DA cell activity is sufficient for the development of cue-induced reward seeking behavior[29,30]. Work in humans using fMRI has implicated striatal targets of DA as learning about state values (ventral striatum) and action policies (dorsal striatum)[31,32], suggesting that dopaminergic signals support both instrumental (action-value) and non-instrumental (state-value) learning in striatum. Consistent with MF value learning, additional research has shown shown that dopaminergic targets, such as dorsal striatum, appear to track MF cached value representations[33,34]. Drug and genetic studies involving humans have shown that variation in dopaminergic function and manipulation of striatal DA sensitivity foster altered learning from positive and negative reward prediction errors[35–37]. Furthermore, DA signals need not be limited to learning outwardly observable 'actions', as projections to cortex have also been suggested to be involved in learning cognitive 'actions' such as determining which items should be held in working memory[36,38–40], implicating the DA learning signal as a general purpose learning signal. In sum, a broad set of methodologies and experimental protocols have shown a consistent link between brain/behavior and computationally defined MF signals associated with the predictive value of the current state (e.e $V(s)$) and/or actions (e.g. $Q(s,a)$) according to motivationally significant events (r_t). Although some work challenges the DA/TD-RL framework[41–43], a broad corpus supports it; the computational RL theory has driven very rich new understanding of learning in the brain.

2.3 Learning as a mixture of MB and MF-RL

More research has built on the successes of using MF RL algorithms to explain brain and behavior by including MB RL as a mechanism through which a broader spectrum of phenomena may be understood. It has long been recognized that animal behavior is not solely determined by reinforcement history, but also exhibits planning characteristics that depend on a cognitive representation of the task at hand[44]. Model-based RL presents a useful computational framework through which this aspect of behavior may be captured.

Attention to MB RL has increased considerably since the creation of the *2-step* task in which the behavioral signatures of MF response and MB planning can be dissociated[7]. In this task, a choice between two available options stochastically leads to one of two second stage states, where a second choice can lead to reward. Each first-level option typically moves the participant into a specific second-stage state (e.g. $a_1 \rightarrow s_1$, and $a_2 \rightarrow s_2$). However, on rare occasions, the participant's choice will lead to the alternative state (e.g. $a_1 \rightarrow s_2$). Choices following rare transitions can dissociate MB from MF RL: MF-RL agents credit reward to the option that was chosen irrespective of the path that led to that reward and will thus be more likely to repeat a rewarded first-stage choice after a rare transition. In contrast, a MB strategy will plan to reach the rewarded second-stage state once more[17], and thus will be less likely to repeat the first-stage choice, favoring the alternative option that most reliably returns it to the reward state (see Figure 2).

Investigations into the relationship between MB/MF-RL and other cognitive/psychological processes have identified links to MB-RL[45–49] more readily than to MF processe[50]. There are several potential explanations for this, one being that the experimental protocols used to probe MB/MF processes, such as the two-step task, are more sensitive to MB control.

Additionally, MB-RL could broadly relate to multiple processes that are highly dependent on a singular mechanism such as attention, offering a unique and easily manipulable cognitive resource through which a disparate processes may be disrupted. Alternatively, this may highlights a problem in the strict dichotomization in learning from MB-MF, as we develop in the next section.

3 Risks

Like any conceptual framework, the MB-MF theory of learning and decision making has intrinsic limitations. Ironically, its increasing popularity and scope of application could erode its potential by advancing a misinterpretation that data must be described along this singular dimension[10]. Indeed, researchers may be led to force a square peg through a round hole when analyzing separable components of their data through the lens of a coarse grained MB-MF dichotomy. Here, we detail some of the more important limitations this presents and how much richer learning theory should become.

3.1 Challenge of disambiguation

3.1.1 MF behaviour can look MB, and vice versa—Despite the ubiquity of MB control[51], labelling behaviour as uniquely MB has been surprisingly difficult[52]. Notably, there are several channels through which behavior rooted in MF cached valuation may emerge to appear reflective of planning, and thus be labeled MB. For example, a MF strategy can flexibly adapt to outcome revaluation in a MB-like way when compound stimuli are formed using previous observations in conjunction with current stimuli[14], a process that has been suggested as a means of transforming a partially observable markov decision processes (POMDPs) into a more tractable MDP[53]. The same can occur when contextual information is used to segregate circumstances in which similar stimuli require different actions[54], or when a model is used retrospectively to identify a previously ambiguous choice[13]. Furthermore, applying a MF learning algorithm to a specific state representation that captures features of trajectories in the environment (e.g. the successor representations[55]), mimics some aspects of MB-behavior (while also making separate predictions). In sum, coupling additional computational machinery such as working memory with standard MF algorithms can mimic a MB planning strategy.

Similarly, there are several paths through which a MB controller may produce behavior that looks MF. For example, one critical measure of MB control is sensitivity to devaluation, where an outcome that had been previously desired is rendered aversive (e.g. by associating that outcome with illness). However, it is not always clear which aspect of MB control has been interfered with should the test subject remain devaluation insensitive (and thus appear MF). In order for MB control to materialize, the agent must first identify its goal, search its model for a path leading to that goal, then act on its plan. Should any of these processes fail (e.g. using the wrong model, neglecting to update the goal, or planning errors), then the agent could appear to be acting more like a MF agent if that is the only alternative under consideration[12,56,57].

Further contributing to the risk of strategy misattribution, non-RL strategies can masquerade as RL when behavior is dichotomized across a singular MB/MF dimension. Simple

strategies that rely only on working memory, such as win-stay/lose-shift, can mimic, or at the very least be difficult to segregate from MF control. Although simple strategies such as WS/LS can be readily identified in tasks explicitly designed to do so[58], more complex non-RL strategies, such as superstitious behavior (e.g. gambler's fallacy, in which losing in the past is though to predict a better chance of winning in the future), or intricate inter-trial patterns (e.g. switch after 2 wins or 4 losses) can be more difficult to identify[59]. Unfortunately, when behavioral response patterns are analyzed within a limited scope along a continuum of being either MB or MF, non-RL strategies are necessarily pressed into the singular axis of MF/MB.

3.1.2 Model use in MF-RL—More generally, other theories of learning assume that agents employ a model of the environment but do not adopt a MB-planning strategy for decision making. For example, the specific type of model used by classic MB algorithms for planning (the transition function) can be used to apply MF-RL updates on retrospectively inferred latent states[13]. This constitutes an example of a class of model-dependent MF-RL algorithms. Models of the environment in this class can include knowledge other than transition and reward functions. A model of the relationship between the outcome of two choices, for example, facilitates counterfactual MF value updates[60,61], while a model of the environment's volatility can be used to dynamically adjust and optimize MF-RL learning rates[62]. Other features of learning using MF-RL updates in conjunction with models of the environment's include work on hidden states, such as non-directly observable rules[54,63–65], demonstrating a rich set of phenomena to which a strict segregation between MB and MF learning and decision making is not well suited.

3.2 MB/MF are not primitive

MB and MF learning are often treated as a singular learning primitives (e.g. "manipulation X increases model-based-ness"). However, the measurable output of either algorithm relies on many computational mechanisms that need not be considered as unique components associated with a singular system. Indeed, MB/MF learning and decision making is arguably better understood as a high-level process that emerges through the coordination of many separable sub-computations, some of which may be shared between the two systems. Thus, the MB/MF dichotomy may not be helpful in identifying unique, separable mechanisms underlying behavior.

3.2.1 Independent underlying computations—It is often forgotten that MB and MF algorithms contain many independent computational sub-components. Although these sub-components are usually thought of from a theoretical perspective as parts uniquely contributing to a particular whole, they may also be recombined in beneficial ways that make the strong separation between MB and MF-RL less meaningful, particularly in light of research investigating their neural implementation and behavioral signatures (fig. 3 B).

For example, MB-RL is characterized by its use of both reward and transition functions to dynamically re-compute expected values. This process, commonly called forward planning, is in fact a high level function that incorporates multiple separable processes. Planning relies on a representation of reward and transition functions; however, it is important to bear in

mind that those representations may not necessarily be used for planning at all[66], or they may serve other processes such as credit assignment, indicating they are not uniquey associated with a "planning" system[13,63]. Furthermore, the transition function, which is often assumed to be known and learned using explicit reasoning[7], may also be shaped using MF-RL-like learning strategy that relies on state prediction errors[67], opening the potential for very different representational structures over which planning must take place. Lastly, planning is simplified by using a mixture of MF and MB valuation whereby MF cached values can be substituted for more costly MB derivations (e.g. by substituting $Q_{MF}(s')$ for $\gamma max_{a'}\,[Q_{MB}(a',s')$ in equation 3 at some point in the planning process[68], suggesting a highly adaptable and varied planning capability. Thus, indicating that *m*anipulation X affects model-based-ness is only weakly informative, as any independent computational sub-component contributing to MB-RL could drive the effect.

Some sub-components may even be shared by the two systems. RL agents make choices by considering scalar values, whether those be dynamically derived (MB) or aggregated cached values (MF). However, agents operating in a real-world environment do not encounter scalar value; rather, they encounter sensory phenomena that must be converted into a valued representation. This translation could be a simple mapping (e.g. a slice of apple is worth 5 units), or it could be conditioned on complex biological and cognitive factors such as the organism's state (hunger, fatigue etc…), the environment (e.g seasonal change, rival competition etc…), or components of the reward itself[69]. Thus, both MF and MB strategies demand some form of reward evaluation process, be it a common resource, or unique to each controller (fig. 3 B).

Similarly, both MB and MF RL algorithms prescribe methods through which option values may be derived, but neither specify how those values should be used to guide decisions (the *policy*). However, the policy has an often important influence on learning: agents need to balance their drive to exploit (pick the best current estimate) and a drive to explore (pick lesser valued options in order to learn more about them). Exploration can be independent of task knowledge (e.g. ε-greedy, where a random choice is made with some probability[22]), or directed toward features of the task model (e.g. uncertainty-guided exploration[70,71]). As such, the action policy, which ultimately guides observable behavior, should be considered independent of the strategy through which valuation, be it MB or MF, occurs.

3.2.2 Independent underlying mechanisms—As we have previously noted, studying brain, behavior, and computational theory through the lens of a MB/MF dichotomy has propelled important advancements across many fields. However, we argue that a singularly dichotomous approach risks promoting an artificial segregation where in fact the computational components that constitute each algorithm are not necessarily unique to either strategy, suggesting they are more richly interconnected than they are distinct. But more importantly to our understanding of brain function and its applicable import (e.g. treatment of mental disease), we suggest that these computations themselves may not map cleanly onto singular underlying neural mechanisms (fig. 3 C). For example, learning a model of the environment and using that model to plan a course of action may rely on shared use of working memory resources[67,72], suggesting some functional overlap at the level of implementation in the brain.

An important, but often overlooked detail is that the primitive functions of RL algorithms assume a pre-defined state and action space[22]. When humans and animals learn, the task space must be discovered, even if MF-RL learning mechanisms then operate over them[54,64,65,73–76]. State space creation likely involves separate networks, such as medial prefrontal cortex[77], lateral prefrontal cortex[74], orbitofrontal cortex[78,79], and hippocampus[80]. Furthermore, a state identification process likely shares functions such as categorization, generalization or causal inference[54,63,64,81]. Critically, the process through which a state space comes to be defined can have dramatic effects on behavioral output. For example, animals can rapidly reinstate response rates following extinction[82,83]. A learning and decision mechanisms that relies on a singular cached value (as is commonly implement using MF-RL) has difficulty capturing this response pattern as it learns, and relearns value symmetrically. However, some implementations can readily elicit reinstatement by learning new representational values for the devalued option, and as such, return to prior response rates rapidly not as a result of learning per se, but as a result of state identification[81,84,85].

Finally, MF value updates may not, in all cases, be a relevant computational primitive matching a clear underlying mechanism to describe behavior, despite the fact that it seems to account for behavioral variance and be reflected in underlying set of neural mechanisms. The family of MF algorithms is extremely broad, and can describe extremely slow learning (such as used to train deep-Q-nets over millions of trials[86], with very low learning rates) or very fast learning (as is often observed in human bandit tasks with high learning rates[87]). It is unlikely that the functions embodied by a singular dopamine-dependent brain network implementing a form of MF-RL are solely responsible for such a broad range of phenomena. Instead it is more likely that the DA-dependent neural MF-RL process is fairly slow (as reflected in the comparably slow learning of many non-human animals), and that faster learning, even when it seemingly can be captured by MF-RL algorithms, actually reflects additional underlying memory mechanisms, such as WM[88–90] and episodic memory[91–95].

In summary, it is important to remember that neither MB nor MF-RL are an atomic unified principal component of learning that map on to unique and separable underlying neural mechanism. The MB-MF dichotomy should be remembered as a convenient description of some aspects of learning that includes forward planning, knowledge of transitions, and outcome valuation, but one that depends on multiple independent sub-components.

3.3 The challenge of isomorphism

The computational MB/MF RL framework has drawn attention as a promising formal lens through which some of the many dichotomous psychological frameworks of decision making may be reinterpreted and unified[11], offering a potential successor to the commonly used but vaguely defined System1/System2 rubric[5,6]. However, hybrid MB/MF RL cannot be the sole basis of a solid theoretical framework for modeling the breadth of learning behavior. In this section, we highlight separable components of learning that do not cleanly align with a MB/MF dichotomization (fig. 3 D), focusing primarily on the habitual vs. goal-directed dichotomy as it is often treated as synonymous with MB and MF RL.[96].

A substantial body of evidence points to two distinguishable modes of behavior: a goal-directed strategy that guides action according to the outcomes they bring about, and habitual

control in which responses are induced by external cues[2]. The principal sources of evidence supporting this dichotomy come from devaluation and contingency degradation protocols aimed at probing outcome directed planning, with the former indexing behavioral adaptations to changes in outcome values, and the latter manipulating the causal relationship between action and outcome (see[97,98] for review). Behavior is considered to be habitual if there's no detectable change in performance despite devalued outcomes or degraded action/outcome contingencies.

The outcome-seeking and stimulus-driven characteristics of goal-directed and habitual behavior mirror the response patterns associated with MB and MF RL respectively[99]. However, as pertinent experimental variables have been probed in more detail, growing evidence suggests that these constructs are not interchangeable. Studies have investigated individual difference measures across the goal-directed/habitual dimension in attempts to relate those to indices of MB/MF control[49,100]. These studies have demonstrated the predicted correspondence between goal-directed response and MB control, but establishing a relationship between habits and MF control has proven more elusive. Indeed, eliciting robust habits is challenging[101], more so than would be expected if habits related to in-lab measures of MF-RL.

Additional facets of learning and decision making have fallen along the emotional axis, with a 'hot' system driving emotionally motivated behavior, and a 'cold' system guiding rational decision making[1,102,103]. Likewise, others have contrasted decisions based on an associative system rooted in similarity based judgements, and a rule based system that guides choice in a logical manner[3,5,6]. Axes have further segregated strategic planning, where one can describe why and how they acted, and implicit "gut-feeling" choice[104,105]. It is tempting to map these contrast to MF/MB RL along a shared *thoughtfulness* axis, but they are theoretically distinct. The MF/MB distinction makes no accommodation for the emotional state of the agent. Both similarity-based judgements and rule creation are beyond most RL algorithms, highlighting fully independent axes of theory, nor has it been established that MB/MF maps cleanly to a contrast between explicit/implicit decision making.

In summary, many dual system frameworks share common themes, thus motivating the more general reference of System 1/System2[5,6]. Although many of the phenomena explained by these dual system frameworks mirror the thrust of the MB/MF dichotomy, none are fully reducible to it. Contrasting some of these dichotomies highlights the fact that MB/MF is not simply a quantitative formalism for those more qualitative theories, but is indeed theoretically distinct from most (e.g. the hot/cold emotional dimension), and offers patchy coverage of others (e.g. habitual/goal-directed).

3.4 What is lost.

Considering other dichotomous frameworks highlights the multi-faceted nature of learning and decision making by showing independent axes along which behavior can be described. Although aligning cognitive/neural/behavioral data across various dualities offers a means by which key variables can be exposed and examined, something is necessarily lost when a system as complex as the brain is scrutinized through a dichotomous lens. Indeed, absorptive terms of description often lack predictive precision (e.g. System1 / System2), while a

proliferation of isolated contrastive frameworks tax our progress toward a coherent understanding of brain and behaviour[106–108]. The application of RL in this campaign marks notable headway by offering a formal framework through which theorems may be proven[109], axiomatic patterns may be described[110], brain function can be probed[29], and theories may be falsified. However, distilling learning and decision making to a single MB/MF dimension risks conflating many other sources of variance, but more importantly, threatens to dilute the formal merits of the framework to that of a verbal theory (e.g. the agent ”uses’ a ”model”).

4 Paths forward

Identifying the computational primitives that support learning is an essential question of basic cognitive (neuro)science, but also has the potential to have important implications in all domains that rely on learning - education, public health, human factors, and so on. It is also of great importance if we are to gain deeper insight into learning differences across populations, including developmental trajectories[111], across environmental factors, or for psychiatric or neurological diseases[112]. Here, we highlight ways in which past research has successfully identified learning primitives that go beyond the MB/MF RL dichotomy, covering many separable dimensions of learning and decision making. These successful approaches offer explicit paths forward in the endeavor of deconstructing learning into its interpretable, neurally implementable basic primitives. This is essential to bridging brain and behavior, and to better understand individual differences across the lifespan as well as in clinical populations[111,112].

Disparities or inconsistencies between classic psychological theoretical frameworks offer opportunities to refine our understanding of the underlying computational primitives. For example, the apparent gaps between MB/MF RL and goal-directed/habitual behavior could promote both theoretical and experimental advances. Failure to elicit a detectable change in post-devaluation response rate using a devaluation protocol (i.e habits) could be caused by a range of mechanisms worthy of further investigation, some of which we have outlined here (e.g. degradation of the transition model, compromised goal maintenance, or engagement of a MF controller etc…). This points to the importance of considering additional dimensions of learning and decision making such as hebbian learning as a mechanism fostering value-free response maintenance[113], and other facets of behavior such as exploration or state-space composition as sources of behavioral variance that may unwittingly appear more MB or MF[14,56].

Computer science research (see 1)also strongly inspires the identification of additional relevant dimensions of learning. For example, algorithms have used hierarchical organization as a means of embedding task abstraction. In hierarchical reinforcement learning (HRL), information is learned and decisions are made at multiple levels of abstraction in parallel. This offers potentially beneficial task abstractions that can span across time[114–116] or the state/action space, and have been observed in humans[54,63,74,117,118]. Notably, HRL may be implemented using either MB planning or MF response, which offers a rich set of computational tools but also compounds the risk of misattribution when a singular MB/MF dimension is considered. Benefit can also come from

considering the classic AI partition between supervised learning, where explicit teaching signals are used to shape system output, and unsupervised learning in which the system relies on properties of the input to drive response. Research has shown that human behavior is shaped by, and exhibits interactions between instructed and experienced trajectories through an environment[39]. Proposals have outlined frameworks where supervised, unsupervised, and RL systems interact in order to build and act on representations of the environment[119,120]; which further bend the notion that a singular spectrum of MB/MF control can sufficiently explain behavior. A third algorithmic dimension that warrants consideration, as it may compound worries of misattribution, is the the distinction between offline and online learning. Online learning agents integrate observations as they arrive, while offline learners can use information at a later point for "batch" updating, relying heavily on information storage and the ability to draw from it[22]. Offline learning has been suggested to occur in between learning trials involving working memory or hippocampal replay[121,122] or during consolidation in sleep[123], and may contribute to both model and reward learning (e.g. the Dyna learning algorithm[22]).

Insights garnered from neuroscience should also continue contributing to enrich our understanding of the dimensions of learning and decision making, as regional specificity has implicated separable aspects of behavior across cortical and subcortical regions. For example, studies in which memory load was systematically manipulated exposed separable roles of MF-RL and working memory in learning[88–90,124], with the two processes mapping on to expected underlying neural systems[88,125,126]. Further examples of using insights from neuroscience to illuminate the computations underlying learning behavior follow from a long history of research into hippocampal function. Previous work has fostered a dichotomy between the hippocampus and the basal ganglia, with the former being implicated in declarative learning, and the latter in implicit procedural learning[127–129]. More recent workhas begun to probe how these two systems may compete for control[91], or collaborate[130]. This collaboration may emerge through relational associations maintained in hippocampus upon which value may be learned[131,132], or through developing a representation that captures transition structure in the environment[133]. Further strengthening a functional relationship, research has also offered evidence of a cooperative computation role between systems during reward learning as a means of actively sampling previous events to improve value estimates[93–95,95].

It is important to note that identifying separable components of learning and decision making is complicated by the existence of interactions between different neural systems. Most theoretical frameworks treat separable components as independent strategies in competition for control. However, they often interact in complex ways beyond competition for choice[134]. For example, in the MB/MF framework[4], striatal signals show that MB information seeps into MF reward prediction error. Similar findings have also been observed in DA recordings[135,136]. Even functions known to stem from largely separable neural underpinning exhibit such interactions: for example information in WM appears to influence MF-RL's RPE computations[89,124–126]. Going beyond simple dichotomies will not only necessitate increasing the dimensionality of the space of learning processes we consider, but also consider how different dimensions interact.

In summary, there are numerous axes along which learning and decision making vary, identified through various traditions of research (e.g psychology, AI and neuroscience). Future research should carry on identifying these axes, and recent work has made much progress identifying many additional dimensions of learning capture other important sources of variance in how we learn, such as meta-learning mechanisms[137,138], learning to use attention[73,139,140], strategic learning[59], and uncertainty-dependent parameter changes[62,141,142]. This is evidence that learning and decision making vary along numerous dimensions that cannot be reduced to a simple two-dimensional principal component space, whether that axis is labelled as MB/MF, hot/cold, goal-directed vs. habitual, or otherwise.

5 Conclusions

We attempted to show the importance of identifying the primitive components supporting learning and decision making as well as the risks inherent to compressing complex and multi-faceted processes into a two-dimensional space. While dual-system theories are a means through which unique and dissociable components of learning and decision-making may be highlighted, key aspects could be fundamentally mis-attributed to unrelated computations, and scientific debate could become counterproductive when different sub-fields use the same label, even as well computationally defined as as MB and MF-RL, to mean different things.

We also propose ways forward. One is to renew a commitment to being precise in our vocabulary and conceptual definitions. The success of the MB-MF RL framework had begun to transition clearly defined computational algorithms toward a range of terms synonymous to many with various dichotomous approximations that may or may not touch on shared functional or neural mechanisms. we have argued that this is a dangerous approximation of a much higher dimensional space. The rigor of computationally defined theories should not hide their limitations: the equations of a model are defined in a precise environment and do not necessarily expand seamlessly to capture neighboring concepts.

Most importantly, we should remember David Marr’s advice and consider our goal when attempting to find primitives of learning[8]. The MB and MF family of algorithms, as defined by computer scientists, offers a high-level theory of what information is incorporated and how it is used during decision making, and how learning is shaped. This may be satisfactory for research that cares about the application of learning science to other domains, such as AI or education. However, for all research whose goal is to understand something that is dependent on the mechanisms of learning (the brain’s implementation), such as the study of individual differences in learning, it is indeed particularly important to ensure that the high-level theory of learning primitives proposes computational primitives that do relate carefully to the underlying circuits. This may benefit from a renewed enthusiasm from computational modelers for the basic building blocks of psychology and neuroscience[143,144], and a better appreciation for the functional atoms formalized by a rich computational theory.

References

1. Roiser JP & Sahakian BJ Hot and cold cognition in depression. CNS spectrums 18, 139–149 (2013). [PubMed: 23481353]

2. Dickinson A Actions and habits: the development of behavioural autonomy. Philos. Transactions Royal Soc. London. B, Biol. Sci 308, 67–78 (1985).
3. Sloman SA The empirical case for two systems of reasoning. Psychol. bulletin 119, 3 (1996).
4. Daw ND, Gershman SJ, Seymour B, Dayan P & Dolan RJ Model-based influences on humans' choices and striatal prediction errors. Neuron 69, 1204–15, DOI: 10.1016/j.neuron.2011.02.027 (2011). [PubMed: 21435563]
5. Stanovich KE & West RF Individual differences in reasoning: Implications for the rationality debate? Behav. brain sciences 23, 645–665 (2000).
6. Kahneman D & Frederick S Representativeness revisited: Attribute substitution in intuitive judgment. Heuristics biases: The psychology intuitive judgment 49, 81 (2002).
7. Daw N Trial-by-trial data analysis using computational models. Decis. making, affect, learning: Atten. …… 1–26 (2011).
8. Marr D & Poggio T A computational theory of human stereo vision. Proc. Royal Soc. London. Ser. B. Biol. Sci 204, 301–328 (1979).
9. Doll BB, Duncan KD, Simon DA, Shohamy D & Daw ND Model-based choices involve prospective neural activity. Nat. neuroscience 18, 767 (2015). [PubMed: 25799041]
10. Daw ND Are we of two minds? Nat. Neurosci 21, 1497–1499, DOI: 10.1038/s41593-018-0258-2 (2018). [PubMed: 30349102]
11. Dayan P Goal-directed control and its antipodes. Neural Networks 22, 213–219 (2009). [PubMed: 19362448]
12. da Silva CF & Hare TA A note on the analysis of two-stage task results: How changes in task structure affect what model-free and model-based strategies predict about the effects of reward and transition on the stay probability. PloS one 13, e0195328 (2018). [PubMed: 29614130]
13. Moran R, Keramati M, Dayan P & Dolan RJ Retrospective model-based inference guides model-free credit assignment. Nat. communications 10, 750 (2019).
14. Akam T, Costa R & Dayan P Simple plans or sophisticated habits? state, transition and learning interactions in the two-step task. PLoS computational biology 11, e1004648 (2015). [PubMed: 26657806]
15. Shahar N et al. Credit assignment to state-independent task representations and its relationship with model-based decision making. Proc. Natl. Acad. Sci 116, 15871–15876 (2019). [PubMed: 31320592]
16. Deserno L & Hauser TU Beyond a cognitive dichotomy: Can multiple decision systems prove useful to distinguish compulsive and impulsive symptom dimensions? Biol. Psychiatry (2020).
17. Doll BB, Duncan KD, Simon D. a., Shohamy D & Daw ND Model-based choices involve prospective neural activity. Nat. neuroscience 1–9, DOI: 10.1038/nn.3981 (2015).
18. Thorndike EL Animal Intelligence: Experimental Studies (Transaction Publishers, 1965).
19. Bush RR & Mosteller F Stochastic models for learning. (1955).
20. Pearce JM & Hall G A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol. review 87, 532 (1980).
21. Rescorla RA, Wagner AR et al. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Class. conditioning II: Curr. research theory 2, 64–99 (1972).
22. Sutton RS & Barto AG Reinforcement learning: An introduction (MIT press, 2018).
23. Montague PR, Dayan P & Sejnowski TJ A framework for mesencephalic dopamine systems based on predictive hebbian learning. J. neuroscience 16, 1936–1947 (1996).
24. Schultz W, Dayan P & Montague PR A Neural Substrate of Prediction and Reward. Science 275, 1593–1599, DOI: 10.1126/science.275.5306.1593 (1997). NIHMS150003. [PubMed: 9054347]
25. Bayer HM & Glimcher PW Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47, 129–41, DOI: 10.1016/j.neuron.2005.05.020 (2005). [PubMed: 15996553]
26. Morris G, Nevet A, Arkadir D, Vaadia E & Bergman H Midbrain dopamine neurons encode decisions for future action. Nat. neuroscience 9, 1057 (2006). [PubMed: 16862149]

27. Roesch MR, Calu DJ & Schoenbaum G Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat. neuroscience 10, 1615 (2007). [PubMed: 18026098]
28. Shen W, Flajolet M, Greengard P & Surmeier DJ Dichotomous dopaminergic control of striatal synaptic plasticity. Science 321, 848–851 (2008). [PubMed: 18687967]
29. Steinberg EE et al. A causal link between prediction errors, dopamine neurons and learning. Nat. neuroscience 16, 966 (2013). [PubMed: 23708143]
30. Kim KM et al. Optogenetic mimicry of the transient activation of dopamine neurons by natural reward is sufficient for operant reinforcement. PloS one 7, e33612 (2012). [PubMed: 22506004]
31. O'Doherty J et al. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. science 304, 452–454 (2004). [PubMed: 15087550]
32. McClure SM, Berns GS & Montague PR Temporal prediction errors in a passive learning task activate human striatum. Neuron 38, 339–346 (2003). [PubMed: 12718866]
33. Samejima K, Ueda Y, Doya K & Kimura M Representation of action-specific reward values in the striatum. Sci. (New York, N.Y.) 310, 1337–40, DOI: 10.1126/science.1115270 (2005).
34. Lau B & Glimcher PW Value representations in the primate striatum during matching behavior. Neuron 58, 451–463 (2008). [PubMed: 18466754]
35. Frank MJ, Seeberger LC & O'Reilly RC By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. Science 306, 1940–1943, DOI: 10.1126/science.1102941 (2004). [PubMed: 15528409]
36. Frank MJ, Moustafa A. a., Haughey HM, Curran T & Hutchison KE Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc. Natl. Acad. Sci. United States Am 104, 16311–6, DOI: 10.1073/pnas.0706111104 (2007).
37. Cockburn J, Collins AG & Frank MJ A reinforcement learning mechanism responsible for the valuation of free choice. Neuron 83, 551–557 (2014). [PubMed: 25066083]
38. Frank MJ, O'Reilly RC & Curran T When memory fails, intuition reigns: midazolam enhances implicit inference in humans. Psychol. science 17, 700–7, DOI: 10.1111/j.1467-9280.2006.01769.x (2006).
39. Doll BB, Hutchison KE & Frank MJ Dopaminergic genes predict individual differences in susceptibility to confirmation bias. J. Neurosci 31, 6188–6198 (2011). [PubMed: 21508242]
40. Doll BB et al. Reduced susceptibility to confirmation bias in schizophrenia. Cogn. Affect. & Behav. Neurosci 14, 715–728 (2014).
41. Berridge KC The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology 191, 391–431 (2007). [PubMed: 17072591]
42. Hamid AA et al. Mesolimbic dopamine signals the value of work. Nat. neuroscience 19, 117 (2016). [PubMed: 26595651]
43. Sharpe MJ et al. Dopamine transients are sufficient and necessary for acquisition of model-based associations. Nat. Neurosci 20, 735 (2017). [PubMed: 28368385]
44. Tolman EC Cognitive maps in rats and men. Psychol. review 55, 189 (1948).
45. Economides M, Kurth-Nelson Z, Lübbert A, Guitart-Masip M & Dolan RJ Model-based reasoning in humans becomes automatic with training. PLoS computational biology 11, e1004463 (2015). [PubMed: 26379239]
46. Otto a. R., Raio CM, Chiang A, Phelps E. a. & Daw ND Working-memory capacity protects model-based learning from stress. Proc. Natl. Acad. Sci. United States Am 110, 20941–6, DOI: 10.1073/pnas.1312011110 (2013).
47. Wunderlich K, Smittenaar P & Dolan RJ Dopamine enhances model-based over model-free choice behavior. Neuron 75, 418–424 (2012). [PubMed: 22884326]
48. Deserno L et al. Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. Proc. Natl. Acad. Sci 112, 1595–1600 (2015). [PubMed: 25605941]
49. Gillan CM, Otto AR, Phelps EA & Daw ND Model-based learning protects against forming habits. Cogn. Affect. & Behav. Neurosci 15, 523–536 (2015).

50. Groman SM, Massi B, Mathias SR, Lee D & Taylor JR Model-free and model-based influences in addiction-related behaviors. Biol. psychiatry 85, 936–945 (2019). [PubMed: 30737015]
51. Doll BB, Simon DA & Daw ND The ubiquity of model-based reinforcement learning. Curr. opinion neurobiology 22, 1075–1081 (2012).
52. Cushman F & Morris A Habitual control of goal selection in humans. Proc. Natl. Acad. Sci 112, 201506367, DOI: 10.1073/pnas.1506367112 (2015). arXiv:1408.1149.
53. O'Reilly RC & Frank MJ Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural computation 18, 283–328 (2006). [PubMed: 16378516]
54. Collins AG & Frank MJ Cognitive control over learning: Creating, clustering, and generalizing task-set structure. Psychol. Rev 120, 190–229, DOI: 10.1037/a0030852 (2013). [PubMed: 23356780]
55. Momennejad I et al. The successor representation in human reinforcement learning. Nat. Hum. Behav 1, 680–692, DOI: 10.1038/s41562-017-0180-8 (2017). [PubMed: 31024137]
56. Da Silva CF & Hare TA Humans are primarily model-based and not model-free learners in the two-stage task. BioRxiv 682922 (2019).
57. Toyama A, Katahira K & Ohira H Biases in estimating the balance between model-free and model-based learning systems due to model misspecification. J. Math. Psychol 91, 88–102 (2019).
58. Iigaya K, Fonseca MS, Murakami M, Mainen ZF & Dayan P An effect of serotonergic stimulation on learning rates for rewards apparent after long intertrial intervals. Nat. communications 9, 1–10 (2018).
59. Mohr H et al. Deterministic response strategies in a trial-and-error learning task. PLoS computational biology 14, e1006621 (2018). [PubMed: 30496285]
60. Hampton AN, Bossaerts P & O'doherty JP The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J. Neurosci 26, 8360–8367 (2006). [PubMed: 16899731]
61. Boorman ED, Behrens TE & Rushworth MF Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. PLoS biology 9, e1001093 (2011). [PubMed: 21738446]
62. Behrens TE, Woolrich MW, Walton ME & Rushworth MF Learning the value of information in an uncertain world. Nat. neuroscience 10, 1214 (2007). [PubMed: 17676057]
63. Collins AGE & Koechlin E Reasoning, learning, and creativity: Frontal lobe function and human decision-making. PLoS Biol. 10, e1001293, DOI: 10.1371/journal.pbio.1001293 (2012). [PubMed: 22479152]
64. Gershman SJ, Norman KA & Niv Y Discovering latent causes in reinforcement learning. Curr. Opin. Behav. Sci 5, 43–50, DOI: 10.1016/j.cobeha.2015.07.007 (2015). NIHMS150003.
65. Badre D, Kayser AS & Esposito MD Article Frontal Cortex and the Discovery of Abstract Action Rules. Neuron 66, 315–326, DOI: 10.1016/j.neuron.2010.03.025 (2010). [PubMed: 20435006]
66. Konovalov A & Krajbich I Mouse tracking reveals structure knowledge in the absence of model-based choice. Nat. communications 11, 1–9 (2020).
67. Gläscher J, Daw N, Dayan P & O'Doherty JP States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66, 585–95, DOI: 10.1016/j.neuron.2010.04.016 (2010). [PubMed: 20510862]
68. Huys QJ et al. Interplay of approximate planning strategies. Proc. Natl. Acad. Sci 112, 3098–3103 (2015). [PubMed: 25675480]
69. Suzuki S, Cross L & O'Doherty JP Elucidating the underlying components of food valuation in the human orbitofrontal cortex. Nat. neuroscience 20, 1780 (2017). [PubMed: 29184201]
70. Badre D, Doll BB, Long NM & Frank MJ Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. Neuron 73, 595–607 (2012). [PubMed: 22325209]
71. Wilson RC, Geana A, White JM, Ludvig EA & Cohen JD Humans use directed and random exploration to solve the explore–exploit dilemma. J. Exp. Psychol. Gen 143, 2074 (2014). [PubMed: 25347535]

72. Otto AR, Gershman SJ, Markman AB & Daw ND The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. Psychol. science 24, 751–61, DOI: 10.1177/0956797612463080 (2013).
73. Niv Y et al. Reinforcement learning in multidimensional environments relies on attention mechanisms. The J. neuroscience : official journal Soc. for Neurosci 35, 8145–57, DOI: 10.1523/JNEUROSCI.2978-14.2015 (2015).
74. Badre D & Frank MJ Mechanisms of Hierarchical Reinforcement Learning in Cortico-Striatal Circuits 2: Evidence from fMRI. Cereb. cortex (New York, N.Y. : 1991) 1–10, DOI: 10.1093/cercor/bhr117 (2011).
75. Collins AGE Reinforcement learning: bringing together computation and cognition. Curr. Opin. Behav. Sci 29, 63–68 (2019).
76. Collins AG Learning structures through reinforcement In Goal-Directed Decision Making, 105–123 (Elsevier, 2018).
77. Donoso M, Collins AGE & Koechlin E Foundations of human reasoning in the prefrontal cortex. Science science. 1252254–, DOI: 10.1126/science.1252254 (2014).
78. Wilson RC, Takahashi YK, Schoenbaum G & Niv Y Orbitofrontal cortex as a cognitive map of task space. Neuron 81, 267–278, DOI: 10.1016/j.neuron.2013.11.005 (2014). NIHMS150003. [PubMed: 24462094]
79. Schuck NW, Wilson R & Niv Y A state representation for reinforcement learning and decision-making in the orbitofrontal cortex In Goal-Directed Decision Making, 259–278 (Elsevier, 2018).
80. Ballard IC, Wagner AD & McClure SM Hippocampal pattern separation supports reinforcement learning. Nat. communications 10, 1073 (2019).
81. Redish AD, Jensen S, Johnson A & Kurth-Nelson Z Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. Psychol. review 114, 784 (2007).
82. Bouton ME Context and behavioral processes in extinction. Learn. & memory 11, 485–494 (2004).
83. Rescorla RA Spontaneous recovery. Learn. & Mem 11, 501–509 (2004).
84. O'Reilly RC, Frank MJ, Hazy TE & Watz B Pvlv: the primary value and learned value pavlovian learning algorithm. Behav. neuroscience 121, 31 (2007).
85. Gershman SJ, Blei DM & Niv Y Context, learning, and extinction. Psychol. review 117, 197 (2010).
86. Wang JX et al. Prefrontal cortex as a meta-reinforcement learning system. Nat. neuroscience 21, 860 (2018). [PubMed: 29760527]
87. Iigaya K et al. Deviation from the matching law reflects an optimal strategy involving learning over multiple timescales. Nat. communications 10, 1466 (2019).
88. Collins AGE & Frank MJ How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. The Eur. journal neuroscience 35, 1024–35, DOI: 10.1111/j.1460-9568.2011.07980.x (2012).
89. Collins AGE The Tortoise and the Hare : Interactions between Reinforcement Learning and Working Memory. J. Cogn. Neurosci 1–12, DOI: 10.1162/jocn (2017).
90. Viejo G, Girard BB, Procyk E & Khamassi M Adaptive coordination of working-memory and reinforcement learning in non-human primates performing a trial-and-error problem solving task. Behav. Brain Res 355, 76–89, DOI: 10.1016/j.bbr.2017.09.030 (2017). 1711.00698. [PubMed: 29061387]
91. Poldrack RA et al. Interactive memory systems in the human brain. Nature 414, 546 (2001). [PubMed: 11734855]
92. Foerde K & Shohamy D Feedback timing modulates brain systems for learning in humans. J. Neurosci 31, 13157–13167 (2011). [PubMed: 21917799]
93. Bornstein AM, Khaw MW, Shohamy D & Daw ND Reminders of past choices bias decisions for reward in humans. Nat. Commun 8, 15958 (2017). [PubMed: 28653668]
94. Bornstein AM & Norman KA Reinstated episodic context guides sampling-based decisions for reward. Nat. neuroscience 20, 997 (2017). [PubMed: 28581478]

95. Vikbladh OM et al. Hippocampal contributions to model-based planning and spatial memory. Neuron 102, 683–693 (2019). [PubMed: 30871859]
96. Decker JH, Otto AR, Daw ND & Hartley CA From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. Psychol. science 27, 848–858 (2016).
97. Dickinson A & Balleine B Motivational control of goal-directed action. Animal Learn. & Behav 22, 1–18 (1994).
98. Balleine BW & Dickinson A Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology 37, 407–419 (1998). [PubMed: 9704982]
99. Daw ND & Doya K The computational neurobiology of learning and reward. Curr. opinion neurobiology 16, 199–204, DOI: 10.1016/j.conb.2006.03.006 (2006).
100. Friedel E et al. Devaluation and sequential decisions: linking goal-directed and model-based behavior. Front. Hum. Neurosci 8, 587, DOI: 10.3389/fnhum.2014.00587 (2014). [PubMed: 25136310]
101. de Wit S et al. Shifting the balance between goals and habits: Five failures in experimental habit induction. J. Exp. Psychol. Gen 147, 1043 (2018). [PubMed: 29975092]
102. Madrigal R Hot vs. cold cognitions and consumers' reactions to sporting event outcomes. J. Consumer Psychol 18, 304–319 (2008).
103. Peterson E & Welsh MC The development of hot and cool executive functions in childhood and adolescence: are we getting warmer? In Handbook of executive functioning, 45–65 (Springer, 2014).
104. Barch DM et al. Explicit and implicit reinforcement learning across the psychosis spectrum. J. abnormal psychology 126, 694 (2017).
105. Taylor JA, Krakauer JW & Ivry RB Explicit and implicit contributions to learning in a sensorimotor adaptation task. J. Neurosci 34, 3023–3032 (2014). [PubMed: 24553942]
106. Sloman SA Two systems of reasoning. (2002).
107. Evans JSBHow many dual-process theories do we need? one, two, or many? (2009).
108. Stanovich K Rationality and the reflective mind (Oxford University Press, 2011).
109. Dayan P The convergence of td (λ) for general λ. Mach. learning 8, 341–362 (1992).
110. Caplin A & Dean M Axiomatic methods, dopamine and reward prediction error. Curr. opinion neurobiology 18, 197–202 (2008).
111. van den Bos W, Bruckner R, Nassar MR, Mata R & Eppinger B Computational neuroscience across the lifespan: Promises and pitfalls. Dev. cognitive neuroscience 33, 42–53 (2018).
112. Adams RA, Huys QJ & Roiser JP Computational psychiatry: towards a mathematically informed understanding of mental illness. J. Neurol. Neurosurg. & Psychiatry 87, 53–63 (2016). [PubMed: 26157034]
113. Miller KJ, Shenhav A & Ludvig EA Habits without values. Psychol. review (2019).
114. Botvinick MM, Niv Y & Barto A Hierarchically organized behavior and its neural foundations: A reinforcement-learning perspective. Cognition 113, 262–280 (2009). [PubMed: 18926527]
115. Konidaris G & Barto AG Skill discovery in continuous reinforcement learning domains using skill chaining. In Advances in neural information processing systems, 1015–1023 (2009).
116. Konidaris G On the necessity of abstraction. Curr. opinion behavioral sciences 29, 1–7 (2019).
117. Frank MJ & Fossella JA Neurogenetics and Pharmacology of Learning, Motivation, and Cognition. Neuropsychopharmacology 36, 133–152, DOI: 10.1038/npp.2010.96 (2010). [PubMed: 20631684]
118. Collins AGE, Cavanagh JF & Frank MJ Human EEG Uncovers Latent Generalizable Rule Structure during Learning. The J. neuroscience 34, 4677–85, DOI: 10.1523/JNEUROSCI.3900-13.2014 (2014).
119. Doya K What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural networks 12, 961–974 (1999). [PubMed: 12662639]
120. Fermin AS et al. Model-based action planning involves cortico-cerebellar and basal ganglia networks. Sci. reports 6, 31378 (2016).

121. Gershman SJ, Markman AB & Otto AR Retrospective revaluation in sequential decision making: A tale of two systems. J. Exp. Psychol. Gen 143, 182 (2014). [PubMed: 23230992]
122. Pfeiffer BE & Foster DJ Hippocampal place-cell sequences depict future paths to remembered goals. Nature 497, 74 (2013). [PubMed: 23594744]
123. Peyrache A, Khamassi M, Benchenane K, Wiener SI & Battaglia FP Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. Nat. neuroscience 12, 919 (2009). [PubMed: 19483687]
124. Collins AGE, Albrecht MA, Waltz JA, Gold JM & Frank MJ Interactions Among Working Memory, Reinforcement Learning, and Effort in Value-Based Choice : A New Paradigm and Selective De fi cits in Schizophrenia. Biol. Psychiatry 82, 431–439, DOI: 10.1016/j.biopsych.2017.05.017 (2017). [PubMed: 28651789]
125. Collins AGE, Ciullo B, Frank MJ & Badre D Working memory load strengthens reward prediction errors. The J. Neurosci 37, 2700–16, DOI: 10.1523/JNEUROSCI.2700-16.2017 (2017).
126. Collins AAGE & Frank MJM Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. Proc. Natl. Acad. Sci. United States Am 115, 1–6, DOI: 10.1073/pnas.1720963115 (2018).
127. Knowlton BJ, Mangels JA & Squire LR A neostriatal habit learning system in humans. Science 273, 1399–1402 (1996). [PubMed: 8703077]
128. Squire LR & Zola SM Structure and function of declarative and nondeclarative memory systems. Proc. Natl. Acad. Sci 93, 13515–13522 (1996). [PubMed: 8942965]
129. Eichenbaum H et al. Memory, amnesia, and the hippocampal system (MIT press, 1993).
130. Foerde K & Shohamy D The role of the basal ganglia in learning and memory: insight from parkinson's disease. Neurobiol. learning memory 96, 624–636 (2011).
131. Wimmer GE, Daw ND & Shohamy D Generalization of value in reinforcement learning by humans. The Eur. journal neuroscience 35, 1092–104, DOI: 10.1111/j.1460-9568.2012.08017.x (2012).
132. Wimmer GE, Braun EK, Daw ND & Shohamy D Episodic Memory Encoding Interferes with Reward Learning and Decreases Striatal Prediction Errors. J. Neurosci 34, 14901–14912, DOI: 10.1523/JNEUROSCI.0204-14.2014 (2014). [PubMed: 25378157]
133. Gershman SJ The successor representation: its computational logic and neural substrates. J. Neurosci 38, 7193–7200 (2018). [PubMed: 30006364]
134. Kool W, Cushman FA & Gershman SJ Competition and cooperation between multiple reinforcement learning systems In Goal-directed decision making, 153–178 (Elsevier, 2018).
135. Langdon AJ, Sharpe MJ, Schoenbaum G & Niv Y Model-based predictions for dopamine. Curr. Opin. Neurobiol 49, 1–7 (2018). [PubMed: 29096115]
136. Starkweather CK, Babayan BM, Uchida N & Gershman SJ Dopamine reward prediction errors reflect hidden-state inference across time. Nat. neuroscience 20, 581 (2017). [PubMed: 28263301]
137. Krueger KA & Dayan P Flexible shaping: How learning in small steps helps. Cognition 110, 380–394 (2009). [PubMed: 19121518]
138. Bhandari A & Badre D Learning and transfer of working memory gating policies. Cognition 172, 89–100 (2018). [PubMed: 29245108]
139. Leong YC et al. Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. Neuron 93, 451–463, DOI: 10.1016/j.neuron.2016.12.040 (2017). [PubMed: 28103483]
140. Farashahi S, Rowe K, Aslami Z, Lee D & Soltani A Feature-based learning improves adaptability without compromising precision. Nat. communications 8, 1768 (2017).
141. Bach DR & Dolan RJ Knowing how much you don't know: a neural organization of uncertainty estimates. Nat. reviews neuroscience 13, 572 (2012). [PubMed: 22781958]
142. Pulcu E & Browning M The misestimation of uncertainty in affective disorders. Trends cognitive sciences (2019).

143. Badre D, Frank MJ & Moore CI Interactionist neuroscience. Neuron 88, 855–860 (2015). [PubMed: 26637794]

144. Krakauer JW, Ghazanfar AA, Gomez-Marin A, MacIver MA & Poeppel D Neuroscience needs behavior: correcting a reductionist bias. Neuron 93, 480–490 (2017). [PubMed: 28182904]

145. Doll BB, Shohamy D & Daw ND Multiple memory systems as substrates for multiple decision systems. Neurobiol. learning memory 117, 4–13, DOI: 10.1016/j.nlm.2014.04.014 (2014). NIHMS150003.

146. Smittenaar P, FitzGerald TH, Romei V, Wright ND & Dolan RJ Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. Neuron 80, 914–919 (2013). [PubMed: 24206669]

147. Doll BB, Bath KG, Daw ND & Frank MJ Variability in dopamine genes dissociates model-based and model-free reinforcement learning. J. Neurosci 36, 1211–1222 (2016). [PubMed: 26818509]

148. Voon V et al. Motivation and value influences in the relative balance of goal-directed and habitual behaviours in obsessive-compulsive disorder. Transl. psychiatry 5, e670 (2015). [PubMed: 26529423]

149. Voon V, Reiter A, Sebold M & Groman S Model-based control in dimensional psychiatry. Biol. psychiatry 82, 391–400 (2017). [PubMed: 28599832]

150. Gillan CM, Kosinski M, Whelan R, Phelps EA & Daw ND Characterizing a psychiatric symptom dimension related to deficits in goaldirected control. eLife 5, 1–24, DOI: 10.7554/eLife.11305 (2016).

151. Culbreth AJ, Westbrook A, Daw ND, Botvinick M & Barch DM Reduced model-based decision-making in schizophrenia. J. abnormal psychology 125, 777 (2016).

152. Patzelt EH, Kool W, Millner AJ & Gershman SJ Incentives boost model-based control across a range of severity on several psychiatric constructs. Biol. psychiatry 85, 425–433 (2019). [PubMed: 30077331]

153. Skinner BF The selection of behavior: The operant behaviorism of BF Skinner: Comments and consequences (CUP Archive, 1988).

154. Corbit LH, Muir JL & Balleine BW Lesions of mediodorsal thalamus and anterior thalamic nuclei produce dissociable effects on instrumental conditioning in rats. Eur. J. Neurosci 18, 1286–1294 (2003). [PubMed: 12956727]

155. Coutureau E & Killcross S Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. Behav. brain research 146, 167–174 (2003).

156. Yin HH, Knowlton BJ & Balleine BW Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur. journal neuroscience 19, 181–189 (2004).

157. Yin HH, Knowlton BJ & Balleine BW Inactivation of dorsolateral striatum enhances sensitivity to changes in the action–outcome contingency in instrumental conditioning. Behav. brain research 166, 189–196 (2006).

158. Ito M & Doya K Distinct neural representation in the dorsolateral, dorsomedial, and ventral parts of the striatum during fixed-and free-choice tasks. J. Neurosci 35, 3499–3514 (2015). [PubMed: 25716849]

Box 1:

Formal RL algorithms

Most commonly, RL problems are formalized as a Markov decision process (MDP), which is defined as; a set of states, S; a set of actions, A; a function $R(s,a)$ that defines the reward delivered after taking action $a \in A$ while in state $s \in S$, and a function $T(s'|s,a)$ that defines which state, $s' \in S$, the agent will transition into if action $a \in A$ is performed while in state $s \in S$

MF-RL algorithms

One approach to solving a reinforcement learning problem is to re-distribute reward information in a way that reflects the environment's structure. MF-RL methods make no attempt to represent the environment's dynamics; rather, they store a set of state/action values that estimate the value of what's expected without explicitly representing the identity of what's to come. This implies that learned values reflect a blend of both the environment's reward and transition structure as encountered reward values are propagated back to be aggregated in preceding states/actions values. For example, having chosen to visit the cafeteria (action a_1) while hungry in their office (state s_1), the student encounters new cafe's booth (state s_2) and samples their food (reward r_1). In one variant of MF-RL, the agent learns about the circumstances that lead to reward using a reward prediction error:

$$\delta = (r_1 + \gamma \cdot Q(a_2, s_2)) - Q(a_1, s_1) \quad (1)$$

$$Q(a_1, s_1) \leftarrow Q(a_1, s_1) + \alpha \cdot \delta \quad (2)$$

where the difference between the predicted value of going to the cafeteria for lunch, $Q(a_1,s_1)$, and the actual value ($r_1 + \gamma\, Q(a_2,s_2)$), is quantified as a temporal difference reward prediction error (δ). The mismatch between expected and experienced outcomes is then used to improve the agent's prediction according to learning rate α (equ 2). Note that both the reward value (r_1) and the discounted expected value of subsequent events ($Q(a_2,s_2)$) are considered as part of the prediction error calculation, offering a path through which rewards can be propagated back to their antecedents.

MB-RL algorithms

As implied by their name, MB algorithms tackle RL problems using a *model* of the environment to plan a course of action by predicting how the environment will respond to its interventions. While *model* can have very different meanings, the model used in MB RL is very specifically defined as the environment's transition function, $T(s'|a,s)$, and reward function, $R(a,s)$. Commonly referenced MB-RL methods either attempt to learn, or are endowed with the model of the task to work with from the start. With a model of the environment, the agent can estimate cumulative state-action values online by planning forward from the current state, or backward from a terminal state. The optimal policy can can be computed using the Bellman equation:

$$Q_{MB}(a_1, s_1) = R(a_1, s_1) + \sum_{s'} T(s' \mid s_1, a_1) \cdot \gamma max_{a'}[Q_{MB}(a', s')] \quad (3)$$

where the value of each action available in the current state, $Q_{MB}(a_1,s_1)$, considers the expected reward $R(a_1,s_1)$, and the discounted expected value of taking the best action at the subsequent state, $\gamma max_{a'}\,[Q(a',s')]$ weighted by the probability of actually transitioning into that state $T(s'/s_1,a_1)$. This approach can be recursively rolled out to subsequent states, deepening the plan under consideration. Thus, when faced with a choice of what to do for lunch, a MB strategy can flexibly consider the value of going back to the cafeteria or of visiting the new cafe by dynamically solving the Bellman equation describing the choice problem.

Box 2:

Learning as a mixture of MB and MF-RL

The original paper reporting the two-step task showed that human behavior exhibited both MB and MF components[7]. Since then, many have used versions of this task to replicate and expand on these findings in what has become rich and productive line of research, highlighting the relevance of MB vs. MF RL in understanding learning across many different domains. We do not provide an exhaustive review here (see[145]), but highlight the impact on neural systems, individual differences, and non-human research to show the breadth of the impact of this theoretical framework on the computational cognitive neuroscience of learning community, and beyond.

Separable neural systems in humans

Subsequent research showed that the dual systems identified by the 2-step task and MB-MF mixture model can be largely mapped to separable systems, either by identifying separate neural correlates[48], or by identifying causal manipulations that taxed the systems independently. Causal manipulations have typically targeted executive functions and as such, the majority (if not all) research using this paradigm have been found to modulate the MB, but not the MF, component of behavior. Successful manipulations that reduced the influence of the MB component included taxing attention via multi-task interference[45] or task-switching[72], inducing stress[46], disrupting regions associated with executive function[146], and pharmacology[47]. Manipulations targeting the MF system are largely absent, potentially pointing to that system's primacy or heterogeneity.

Individual differences

Individuals vary in their decision making process and how they learn from feedback. The MB-MF theoretical framework, along with the 2-step task, was successfully used to capture such individual differences and relate them to predictive factors[147]. For example, in a developmental cohort,[96] showed that the MB component increased from age 8 through 25, while the MF component of learning remained stable. This framework has also been used to identify specific learning deficits in psychiatric populations, such as people with obsessive compulsive disorders[148] or repetitive disorders[149],addiction[150], schizophrenia[151] and other psychiatric constructs[49,152].

Non-human studies

Early models of animal behavior described a causal relationship between stimuli and response[153], which was expanded upon to show that some behavior was better accounted for by models that included a cognitive map of the environment[44]. However, more refined investigations suggested that both strategies, a stimulus-driven response and an outcome motivated action, can emerge from the same animals[2]. Anatomical work in rats has dissociated these strategies, indicating that pre-limbic regions are involved in goal-directed learning[98,154], while infralimbic cortex has been associated with S-R control[155]. This dissociation mirrors a functional segregation between dorsolateral and dorsomedial striatum, with the former implicated in S-R behavior, and the later being associated with goal-directed planning[156–158].

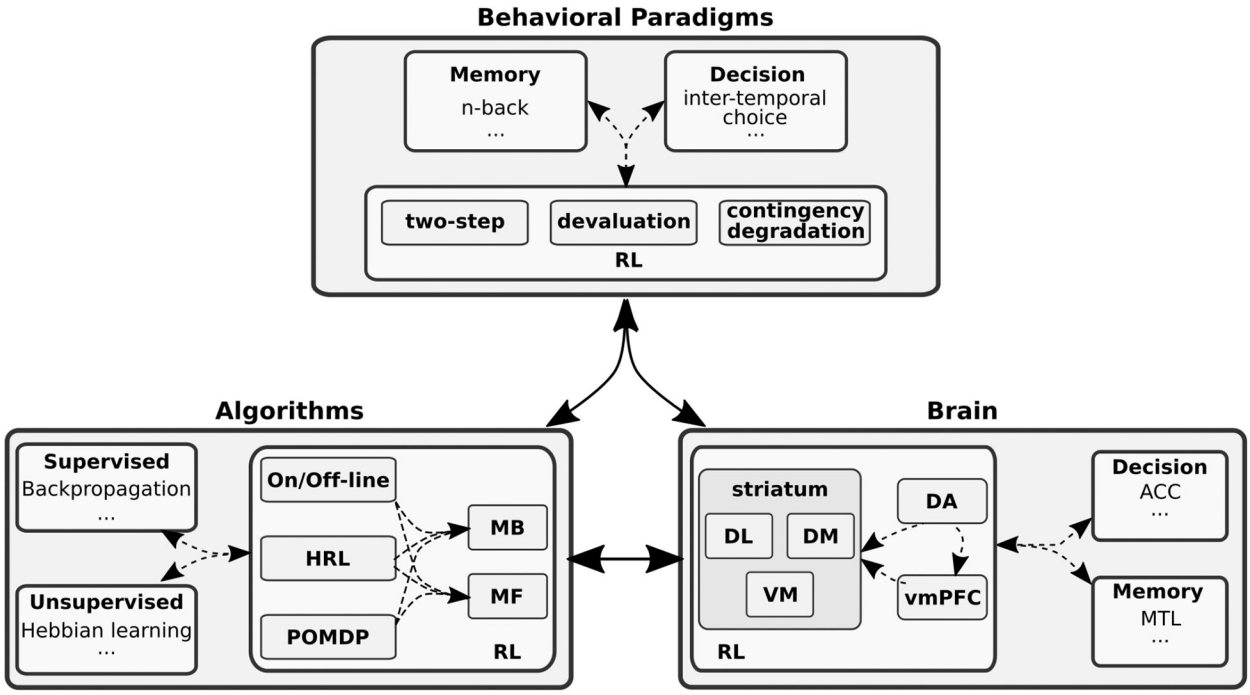

Figure 1. RL across fields of research

Many fields of research use the term reinforcement learning (RL), notably computational, behavioral, and neurosciences. The meaning of RL in each field is used in contrast to other concepts (e.g. supervised in machine learning). While computational sciences frames dichotomies between algorithmic approaches, behavioral sciences contrast and define cognitive constructs by way of experimental designs (e.g. habits are devaluation insensitive behaviors[2]), and neuroscience focuses on the brain's separable neural circuits. It is also well accepted that the segregation, both conceptually and empirically, are practical though imperfect simplifications. For example, both memory and decision making processes make significant contributions to the neural circuits involved in RL, meaning that brain regions not uniquely associated with RL contribute to RL behavior nonetheless (dashed arrows). It is important to remember that while the three RL definitions are related (full arrows), they are not equivalent. dorso-lateral (DL); dorso-medial (DM); ventral-medial (VM); ventral-medial prefrontal cortex (vmPFC); dopamine (DA); anterior cingulate cortex (ACC); mediotemporal lobe (MTL).

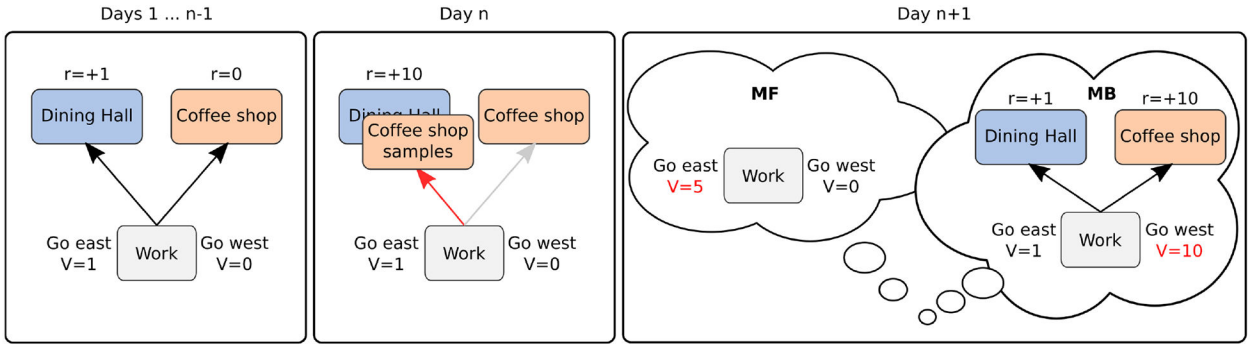

Figure 2. Contrast between MB and MF algorithms in response to environmental changes.

A student has learned that the cafeteria is to the east of their lab, and the coffee shop is to the west. Having visited both several times in the past, they have also learned that the lunch offerings at the cafeteria are passable (reward of +1), while the coffee shop does not offer food (reward of 0). On day *n*, the student opts to visit the cafeteria (which both MB and MF strategies agree to as the best option). However, the student encounters a stand in front of the cafeteria offering delicious items from a new menu at the coffee shop (reward of +10). The next day, the student must decide which direction to take for lunch. A MB strategy will consult its model of the environment to identify the path toward the best lunch option, which is now at the coffee shop (go west). A MF strategy, in contrast, will consult its value estimates, and owing to the unexpectedly good lunch the previous day, it will repeat the action of heading east (toward the cafeteria).

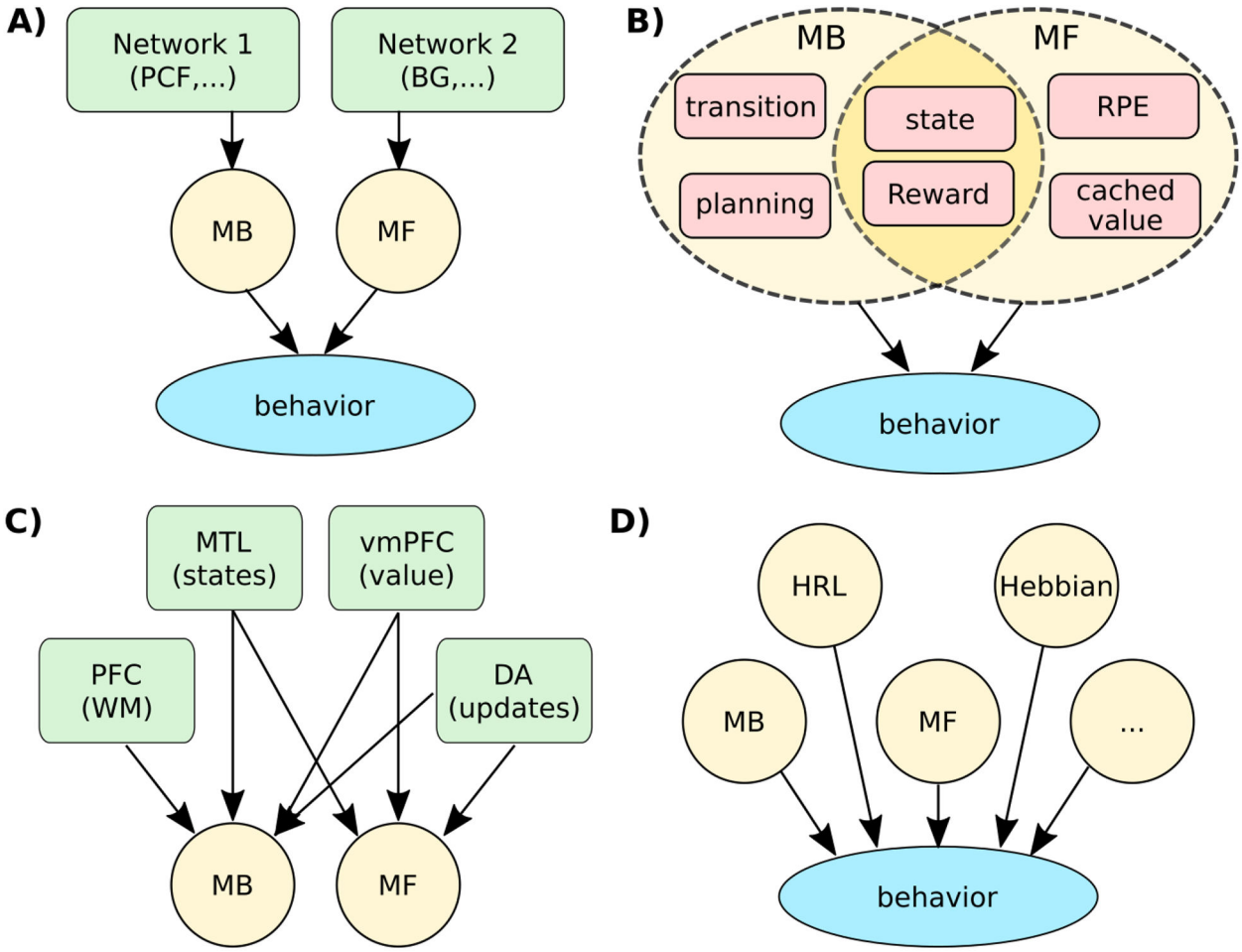

Figure 3.

Decompositions of learning. **A**. Classic interpretations of the MB-MF RL theory cast the space of learning behavior as a mixture of two components, with MB and MF as independent primitives implemented in separable neural networks (green). **B**) In reality, MB and MF are not independent computational dimensions, and rely on multiple partially shared computational primitives (red). For example, MB planning depends on learned transitions, which in turn, relies on state representations that may be shared across MB/MF strategies. **C**) MB and MF's computations do not map on to unique underlying mechanisms. For example, MB learning may rely on prefrontal (PFC) working memory to compute forward plans, medial temporal lobe (MTL) to represent states and transition, and ventro-medial (vm) PFC to represent reward expectations. MF also relies on the latter two, as well as other specific networks, non-exhaustively represented here. **D**) Additional independent computational dimensions are needed to account for the space of learning algorithm behaviors, such as hierarchical task decomposition (HRL) or hebbian learning.