Beyond dichotomies in reinforcement learning

Anne G. E. Collins and Jeffrey Cockburn

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 that relate multiple levels of analysis in a singular framework (for example, relating dopamine function to a computationally defined RL signal). Recently, more sophisticated RL algorithms have been proposed to better account for human learning, and in particular its oft-documented reliance on two separable systems: a model-based (MB) system and a model-free (MF) system. 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 overconfidently mapping algorithms, such as MB versus MF RL, with putative cognitive processes. We argue that the field is well positioned to move beyond simplistic dichotomies, and we propose a means of refocusing research questions towards the rich and complex components that comprise learning and decision-making.

The empirical study of learning and decision-making, in both humans and non-human animals, has catalogued a wealth of evidence consistent with the idea that behaviour is governed by at least two separable controllers. Behaviour has been dichotomized across several dimensions, including emotion (hot–cold)[1], action selection (habitual–goal-directed)[2], judgements (associative–rule-based)[3] and, more recently, reinforcement learning (RL; model-free (MF)–model-based (MB))[4]. The terms used to characterize these controllers vary, but have largely been absorbed into the terms System1/System2 (REFS[5,6]). Thus, whereas many seemingly 'irrational' behaviours have been argued to emerge from a system that is fast, reactive, implicit, retrospective and emotionally charged, another system has been posited to support behaviours that are described as slow, deliberative, explicit, prospective and calculated[5,6]. Our understanding of the processes that drive behaviour, from the neural implementations to social factors, has advanced considerably through the use of these dichotomies in terms of both experimental and theoretical development.

However, despite a common philosophical core, the various frameworks used to describe these behavioural controllers vary in terms of their formalism and scope and, as such, neither they nor the phenomena they purport to explain are interchangeable. 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 behaviour as being driven by any one particular controller. To address this, dual-system theories of learning and decision-making have been drawn towards the formalization offered by the field of machine learning, readily found in the literature as a mapping to MB or MF RL[7].

Computational formalization promises important benefits: it promotes a precise quantitative definition of key concepts, and often enables 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. Although 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 processes by mapping them onto two a-priori defined components may cause the field to lose sight of their essential features. We focus on the example of the MB–MF RL dichotomy for one key reason: it 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[11]. We show that this confidence, which is 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 behaviour 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 behaviour may offer a more fruitful path forward.

What is reinforcement learning?

RL is a term widely used in at least three separate, but overlapping, fields of research: computational sciences (including machine learning, artificial intelligence (AI) and computer science); behavioural sciences (psychology and cognitive science); and neuroscience (including systems and cellular neuroscience) (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). Behavioural sciences use RL to refer to learning processes that promote behaviour by pairing it with a valued outcome (or the removal of an undesired outcome), and discourage otherwise. The field of neuroscience typically describes RL as a form of dopamine (DA)-dependent plasticity that shapes neural pathways to implement value-based learning in the brain (specifically corticostriatal networks).

RL algorithms

Computational RL defines a class of learning problems and algorithms, such as MF and 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 (for example, playing tic-tac-toe) 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. However, most real-world problems (such as driving to work) are far more complex: the number of possible circumstances in which the agent might find itself (that is, the state space) can be huge, as are the number of available actions, and measures of 'progress' can be difficult to quantify. 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) the 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, they derive value estimates for the different states or actions available; then, they choose actions that are deemed most valuable.

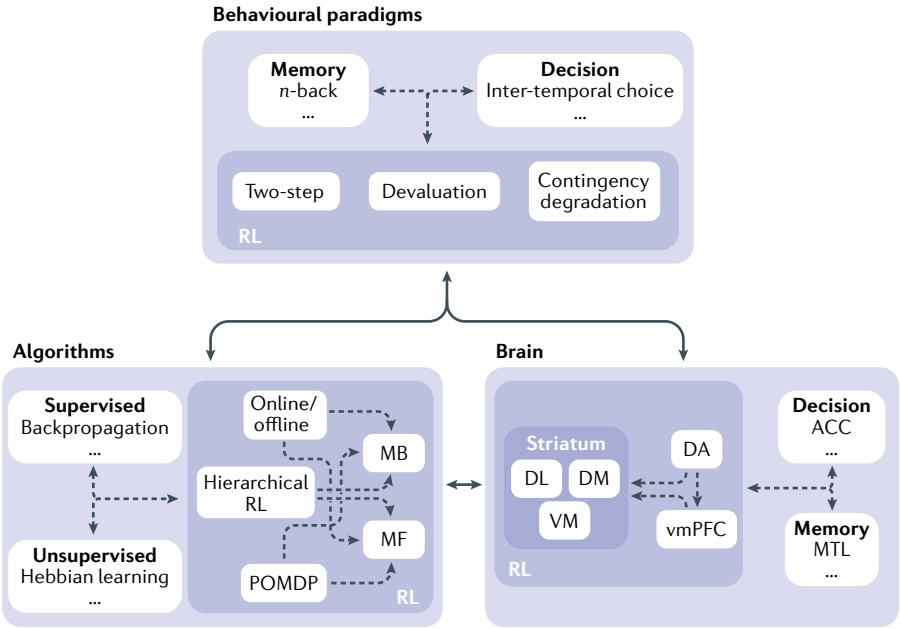

Fig. 1 | **RL across fields of research.** Many fields of research use the term reinforcement learning (RL), notably computational science, behavioural science and neuroscience. The meaning of RL in each field is used in contrast to other concepts (for example, supervised learning within machine learning). Whereas computational science frames dichotomies between algorithmic approaches, behavioural science contrasts and defines cognitive constructs by way of experimental designs (for example, habits are devaluation-insensitive behaviours[2]) and neuroscience focuses on the brain's separable neural circuits. It is also well accepted that the segregation, both conceptually and empirically, is a practical yet imperfect simplification. For example, both memory and decision-making processes make substantial contributions to RL, meaning that brain regions not uniquely associated with RL nevertheless contribute to RL behaviour (dashed arrows). It is important to remember that although the three RL definitions are related (full arrows), they are not equivalent. ACC, anterior cingulate cortex; DA, dopamine; DL, dorsolateral striatum; DM, dorsomedial striatum; MB, model-based; MF, model-free; MTL, mediotemporal lobe; POMDP, partially observable Markov decision process; VM, ventromedial striatum; vmPFC, ventromedial prefrontal cortex.

RL algorithms can be categorized along many dimensions. MB and 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 including the transition and reward functions. 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 towards its goal. 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 (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 towards an arbitrary goal; however, they suffer the additional hindrance of computing this action plan, which can rapidly become intractable. MF algorithms cannot adapt as easily, owing to their strategy of integrating reward history into a single value estimate; however, they offer an efficient approach to learning and decision-making.

As an example, consider 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. By contrast, a MF strategy would consult its value estimates and simply repeat yesterday's choice to visit the cafeteria, as 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, the MF choice is considerably less effortful, as it relies on

a cached value estimate that can be derived using simple computations that rely only on easily accessible information (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[9] 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[4,17,18]. RL has been pivotal in providing a sound quantitative theory of learning, and a normative framework through which we can understand the brain and behaviour. 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.

MF RL and the brain

Early research into the principles that govern learning likened behaviour to the output of a stimulus–response association machine that forms links between stimuli and motor responses through reinforcement[19]. Various models described the relationships between stimuli, response and reward, with nearly all sharing a common theme of an associative process driven by a surprise signal[20–22]. Computational RL theory built on the principles that animal behaviourists had distilled through experimentation, to develop the method of temporal difference (TD) learning (a MF algorithm), which offers general-purpose learning rules while also formalizing the RL problem[23].

The TD RL algorithm sparked a turning point in our understanding of DA function in the brain. In a seminal set of studies, the phasic firing patterns of DA neurons in the ventral tegmental area were shown to mirror the characteristics of a TD RL reward prediction error (see Eq. (1) in BOX 1), offering a bridge between behaviourally descriptive models and a functional understanding of learning algorithms embodied by the brain[17,24,25]. Continued work along this line of research has probed the details of DA activity in greater detail, linking it to various flavours of MF RL[26,27]. Importantly, this work has shifted the conceptualization of stimulus–response instrumental learning away from inflexible reflex-like behaviour towards one of adaptable, value-based learning.

The role of DA as a MF RL teaching signal is supported by work in both humans and non-human animals showing that DA affects corticostriatal plasticity, as theoretically predicted[28]. Subsequent research has focused on the causal importance of DAergic input to show that systematic modulation of DA cell activity is sufficient for the development of cue-induced reward-seeking behaviour[29,30]. Work in humans using functional MRI has implicated the ventral and dorsal striatal targets of DA in learning about state values and learning about action policies, respectively[31,32], suggesting that DAergic signals support both instrumental (action–value) and non-instrumental (state–value) learning in the striatum. Consistent with MF value learning, additional research has shown that DAergic targets such as the dorsal striatum seem to track MF cached value representations[33,34]. Pharmacological and genetic studies involving humans have shown that variation in DAergic function and the manipulation of striatal DA sensitivity foster altered learning from positive and negative reward prediction errors[35–37]. Furthermore, DA signals need

Box 1 | **Formal RL algorithms**

Most commonly, reinforcement learning (RL) problems are formalized as a Markov decision process, 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 $s \in S$.

Model-free RL algorithms

One approach to solving a RL problem is to redistribute reward information in a way that reflects the environment's structure. Model-free (MF) RL methods make no attempt to represent the dynamics of the environment; rather, they store a set of state or action values that estimate the value of what is expected without explicitly representing the identity of what is to come. This implies that learned values reflect a blend of both the reward structure and the transition structure of the environment, as encountered reward values are propagated back to be aggregated with preceding state values or action values. For example, having chosen to visit the cafeteria (action a_1) while hungry in their office (state s_1), a student encounters the 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 led to reward using a reward prediction error. Specifically, the difference between the predicted value of going to the cafeteria for lunch, $Q(a_1, s_1)$, and the actual value, $r_1 + \gamma \cdot Q(a_2, s_2)$, where γ discounts future value relative to immediate reward, is quantified as a temporal difference reward prediction error (δ):

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

The mismatch between the expected outcome and the experienced outcome is then used to improve the agent's prediction according to learning rate α:

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

Note that both the reward value (r_1) and the discounted expected value of subsequent events ($\gamma \cdot Q(a_2, s_2)$) are considered as part of the prediction error calculation, offering a path through which rewards can be propagated back to their antecedents.

Model-based RL algorithms

As implied by their name, model-based (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. Although the word 'model' can have very different meanings, the model used in MB RL is very specifically defined as the transition function, $T(s'|a, s)$, and the reward function, $R(a, s)$, of the environment. Commonly referenced MB RL methods either attempt to learn, or are endowed with, the model of the task. 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 be computed using the Bellman equation, in which the value of each action available in the current state, $Q_{MB}(a_1, s_1)$, takes into account the expected reward $R(a_1, s_1)$, and the discounted expected value of taking the best action at the subsequent state, $\gamma \cdot \max_{a'}[Q(a', s')]$, weighted by the probability of actually transitioning into that state $T(s'|s_1, a_1)$:

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

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.

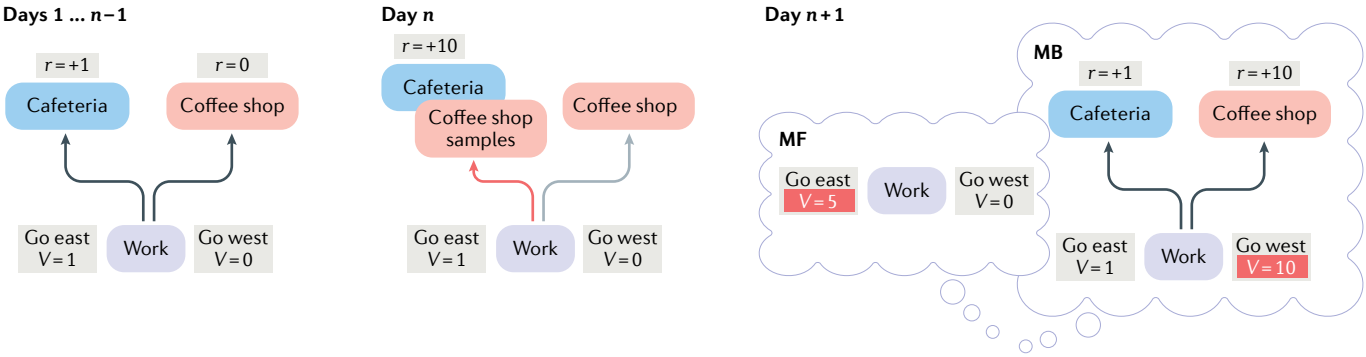

Fig. 2 | **Contrast between MB and MF algorithms in response to environmental changes.** (Left) A student has learned that the cafeteria is to the east of their laboratory and the coffee shop is to the west. Having visited both several times in the past, they have also learned that the lunch offerings at the cafeteria are passable (reward (r) = +1), whereas the coffee shop does not offer food (r = 0). (Middle) On day n, the student opts to visit the cafeteria (with value V(east) = 1, and V(west) = 0, both model-based (MB) and model-free (MF) strategies agree going east to the cafeteria is 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 (r = +10). (Right) 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 towards the best lunch option, which is now at the coffee shop (go west). A MF strategy, by contrast, will consult its value estimates and, owing to the unexpectedly good lunch the previous day, will repeat the action of heading east (towards the cafeteria).

not be limited to learning outwardly observable 'actions', as projections to the 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], thus implicating the DA learning signal as a general-purpose learning signal. In sum, a broad set of methodologies and experimental protocols have shown consistent links between brain, behaviour and computationally defined MF signals associated with the predictive value of the current state and/or actions according to motivationally salient events such as reward. Although some work challenges the DA-dependent TD RL framework[41–43], a broad corpus supports it, and the computational RL theory has driven very rich, new understanding of learning in the brain.

A mixture of MB and MF RL

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

Attention to MB RL has increased considerably since the creation of the two-step task, in which the behavioural signatures of MF responses 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, at which a second choice can lead to reward. Each first-level option typically moves the participant into a specific second-stage state (such as $a_1 \rightarrow s_1$ and $a_2 \rightarrow s_2$). However, on rare occasions, the participant's choice will lead to the alternative state (for example, $a_1 \rightarrow s_2$). Choices following rare transitions can dissociate MB RL from MF RL: MF RL agents credit reward for 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. By contrast, a MB strategy will plan to reach the rewarded second-stage state once more[9], and thus will be less likely to repeat the first-stage choice, favouring the alternative option that most reliably returns it to the reward state (FIG. 2).

Investigations into the relationship between MB and MF RL and other cognitive or psychological processes have identified links with MB RL[45–49] more readily than with MF processes[50]. There are several potential explanations for this, one being that the experimental protocols used to probe MB and MF processes, such as the two-step task, are more sensitive to MB control. In addition, MB RL could broadly relate to multiple processes that are highly dependent on a single mechanism, such as attention, which offers an easily manipulable channel through which many subservient processes may be disrupted. Alternatively, the imbalance tilted in favour of MB cohesion across theoretical boundaries may highlight a problem in the strict dichotomization in learning from MB–MF, as we develop in the next section.

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 analysing 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.

Challenge of disambiguation

MF behaviour can look MB, and vice versa. Despite the apparent ubiquity of MB control in human behaviour[51], labelling behaviour as uniquely MB has been surprisingly difficult[52]. Notably, there are several channels through which behaviour that depends on a MF cached valuation may seem to reflect planning, and thus be labelled MB. For example, a MF strategy can flexibly adapt in a MB-like way when learners form compound representations using previously observed stimuli and outcomes in conjunction with current stimuli[14], a process that has been offered as a means of transforming a partially observable Markov decision process (where task dynamics are determined by the current state of the environment, but the agent cannot directly observe the underlying state) into a more

Box 2 | **Learning as a mixture of MB and MF RL**

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

Separable neural systems in humans

The dual systems identified by the two-step task and the MB–MF mixture model were shown to largely map 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) of research using this paradigm has been found to modulate the MB, but not the MF, component of behaviour. Successful manipulations that reduced the influence of the MB component included taxing attention via multitask interference[45] or task-switching[72], inducing stress[46], disrupting regions associated with executive function[146] and pharmacological treatments[47]. Manipulations targeting the MF system are largely absent, potentially reflecting that system's primacy or heterogeneity.

Individual differences

Individuals vary in their decision-making processes and how they learn from feedback. The MB–MF theoretical framework, along with the two-step task, was successfully used to capture such individual differences and relate them to predictive factors[147]. For example, a study of a developmental cohort[96] showed that the MB component increases from age 8 through 25 years, whereas the MF component of learning remains 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 behaviour described a causal relationship between stimuli and responses[153], which was expanded upon to show that some behaviour 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 prelimbic regions are involved in goal-directed learning[98,154], whereas the infralimbic cortex has been associated with stimulus–response control[155]. This dissociation mirrors a functional segregation between the dorsolateral and dorsomedial striatum, with the former implicated in stimulus–response behaviour and the latter being associated with goal-directed planning[156–158].

tractable Markov decision process, where the state is fully observed[53]. In similar fashion, MB-like behaviour can emerge from a MF controller 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 representations that capture features of trajectories in the environment (for example, successor representations that track the frequency with which the agent arrives at a given state[55]) mimics some aspects of MB behaviour (yet makes 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 behaviour that looks MF. For example, one important indication of MB control is sensitivity to devaluation, whereby an outcome that had been previously desired is rendered aversive (for example, through association with illness). However, it is not always clear which aspect of MB control has been disrupted if the agent remains devaluation-insensitive (and, thus, seems MF). For MB control to materialize, the agent must identify its goal, search its model for a path leading to that goal and then act on its plan. Should any of these processes fail (for example, using the wrong model, neglecting to update the goal or planning errors), then the agent could seem to act 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 behaviour 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 distinguish from — MF control. Although simple strategies such as 'win–stay/lose–shift' can be readily identified in tasks explicitly designed to do so[58], more complex non-RL strategies, such as superstitious behaviour (for example, gambler's fallacy, in which losing in the past is believed to predict a better chance of winning in the future) or intricate inter-trial patterns of responding (for instance, switch after two wins or four losses), can be more difficult to identify[59]. Unfortunately, when behavioural response patterns are analysed 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.

Model use in MF RL. More generally, other theories of learning assume that agents use 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 also be used to apply MF RL updates on retrospectively inferred latent states[13]. This combination of a MB feature in otherwise MF learning 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 functions and reward functions. For example, a model of the relationship between the outcome of two choices facilitates counterfactual MF value updates[60,61], whereas a model of the environment's volatility can be used to dynamically adjust and optimize MF RL learning rates[62]. Learning using MF RL updates in conjunction with models of the environment also occurs in the context of identifying hidden states, such as non-directly observable rules[54,63–65]. MF learning with model use is thus involved in a rich set of learning experiences, meaning that a strict segregation between MB and MF learning and decision-making is not easily justified or helpful.

MB and MF learning are not primitive

MB and MF learning are often treated as a singular learning primitive (for example, 'manipulation X increases MB-ness'). However, the measurable output of either a MF or a MB algorithm relies on many computational mechanisms that need not be considered as unique components associated with a singular system. Indeed, MB and 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 behaviour.

Independent underlying computations. It is often forgotten that MB and MF algorithms contain many independent computational subcomponents. Although these subcomponents are usually considered from a theoretical perspective as parts that make different contributions 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 behavioural signatures (FIG. 3b).

For example, MB RL is characterized by its use of reward functions and transition functions to dynamically recompute 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 functions and transition functions; however, such representations may not necessarily be used for planning at all[66], or may serve other processes such as credit assignment, indicating they are not uniquely associated with a 'planning' system per se[13,63]. Furthermore, the transition function, which is often assumed to be known and learned using explicit reasoning[7], may also be shaped through a MF RL-like learning strategy that quantifies the discrepancy between the expectation of a state transition and what is actually observed[67]. This opens the potential for very different representational structures — those shaped by experience and those shaped by explicit instruction — over which planning must take place. Last, planning is simplified by using a mixture of MF and MB valuation, whereby MF cached values can be substituted for more costly MB derivations (for example, by substituting $Q_{MF}(s')$ for $\gamma \max_{a'}[Q_{MB}(a', s')]$ in Eq. 3 in BOX 1) at some point in the planning process[68], suggesting that planning ability can be highly adaptable and varied. Thus, indicating that manipulation X affects MB-ness is only weakly informative, as any independent computational subcomponent contributing to MB RL could drive the effect of manipulation X.

Some subcomponents may even be shared by the two systems. RL agents make choices by considering scalar values, whether dynamically derived (MB) or aggregated and cached (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 (for example, a slice of apple is worth 5 units), or it could be conditioned on complex biological and cognitive factors such as those relating to the organism's state (hunger, fatigue and so on), the environment (such as seasonal change, rival competition and so on) or components of the reward itself (such as vitamin, carbohydrate and fat levels)[69]. Thus, both MF and MB strategies demand some form of reward-evaluation process, whether this process is common or specific to each of these strategies (FIG. 3b).

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 (that is, the 'policy'). However, the policy has an often important influence on learning: agents need to balance their drive to exploit (by picking the best current estimate) and a drive to explore (by picking lesser-valued options in order to learn more about them). Exploration can be independent of task knowledge (for example, in an ε-greedy strategy, in which a random choice is made with some probability[23]) or directed towards features of the task model (for example, guided by uncertainty[70,71]). As such, the action policy, which ultimately guides observable behaviour, should be considered independent of the strategy through which valuation, be it MB or MF, occurs.

Independent underlying mechanisms. As we have previously noted, studying brain, behaviour 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 that

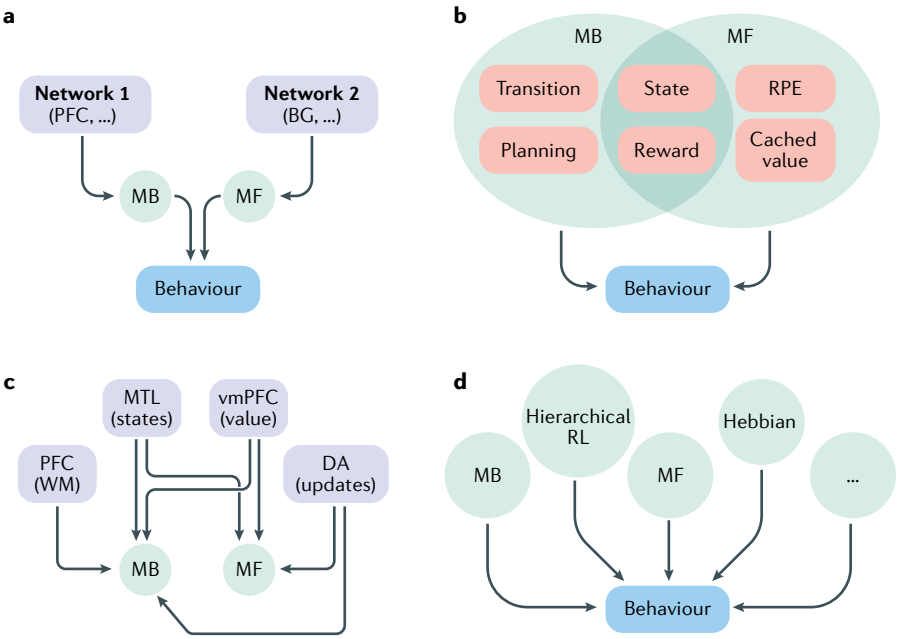

Fig. 3 | **Decompositions of learning. a** | Classic interpretations of the model-based (MB)–model-free (MF) reinforcement learning (RL) theory cast the space of learning behaviour as a mixture of two components, with MB and MF as independent primitives implemented in separable neural networks (green). **b** | In reality, MB and MF RL are not independent computational dimensions, and rely on multiple partially shared computational primitives (darker green overlap). For example, MB planning depends on learned transitions, which in turn, relies on state representations that may be shared across MB and MF RL strategies. **c** | The computations supportive of MB and MF RL do not map to unique underlying mechanisms. For example, MB learning may rely on working memory (WM) in the prefrontal cortex (PFC) to compute forward plans, the medial temporal lobe (MTL) to represent states and transition, and the ventromedial PFC (vmPFC) to represent reward expectations. MF RL also relies on the latter two, as well as other specific networks, non-exhaustively represented here. **d** | Additional independent computational dimensions, such as hierarchical task decomposition (hierarchical RL) or Hebbian learning, are needed to account for the range of learning algorithm behaviours. BG, basal ganglia; DA, dopamine; RPE, reward prediction error.

the cognitive processes are more richly interconnected than they are distinct. But more importantly for our understanding of brain function and its applicable import (for example, in the treatment of mental disease), we suggest that these computations themselves may not map cleanly onto singular underlying neural mechanisms (FIG. 3c). For example, learning a model of the environment, as contrasted with using that model to plan a course of action, may rely on a common 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 predefined state and action space[23]. 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]. Shaping a state space to represent the relevant components of the environment for the task at hand probably involves separate networks, such as the medial prefrontal cortex[77], lateral prefrontal cortex[74], orbitofrontal cortex[78,79] and hippocampus[80]. Furthermore, a state-identification process probably depends on complex, multipurpose 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 behavioural output. For example, animals can rapidly reinstate response rates following extinction[82,83]. A learning and decision mechanism that relies on a singular cached value (as is commonly implement in MF RL) has difficulty capturing this response pattern as, in MF, value is learned and relearned symmetrically. However, some implementations of MF RL 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 representative of a unique underlying mechanism, despite the fact that it seems to account for behavioural variance and be reflected in a set of underlying neural mechanisms. The family of MF algorithms is extremely broad, and can underlie extremely slow learning (as is used to train deep Q-nets over millions of trials[86]) or very fast learning (as often observed in human bandit tasks with high learning rates[87]). It is unlikely that the functions embodied by a singular DA-dependent brain network that implements 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 working memory[88–90] and/or episodic memory[91–95].

In summary, it is important to remember that MB RL and MF RL are not atomic principal components of learning and decision-making that map to unique and separable underlying neural mechanisms. The MB–MF dichotomy should be considered a convenient description of some aspects of learning, including forward planning, knowledge of transitions and outcome valuation, but one that depends on multiple independent subcomponents.

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 modelling the breadth of learning behaviour. In this section, we highlight separable components of learning that do not cleanly align with a MB–MF dichotomization (FIG. 3d), focusing primarily on the habitual versus goal-directed dichotomy, as this is often treated as synonymous with MB and MF RL[96].

A substantial body of evidence points to two distinguishable modes of behaviour: a goal-directed strategy that guides actions according to the outcomes they bring about; and habitual control, through 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 behavioural adaptations to changes in outcome values, and the latter manipulating the causal relationship between action and outcome (see REFS[97,98] for a review). Behaviour is considered habitual if there is 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 behaviour 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 differences in 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 proved more elusive. Indeed, eliciting robust habits is challenging[101], more so than would be expected if habits related to in-laboratory measures of MF RL.

Additional facets of learning and decision-making have fallen along the emotional axis, with a 'hot' system driving emotionally motivated behaviour and a 'cold' system guiding rational decision-making[1,102,103]. Similarly, others have contrasted decisions based on an associative system rooted in similarity-based judgements with decisions derived from a rule-based system that guides choice in a logical manner[3,5,6]. Studies and theory have further segregated strategic planning, whereby one can describe why and how they acted, and implicit 'gut-feeling' choice[104,105]. Although it is tempting to map these various dichotomies to MF–MB RL along something akin to a common 'thoughtfulness' dimension, they are theoretically distinct. The MF–MB distinction makes no accommodation for the emotional state of the agent. Similarity-based judgements and rule creation are not generally addressed by RL algorithms, and the MB–MF dichotomy has not been cleanly mapped to a contrast between explicit and implicit decision-making.

In summary, many dual-system frameworks share common themes, thus motivating the more general reference of System1/System2 (REFS[5,6]). Although many of the phenomena explained by these dual-system frameworks mirror the gist of the MB–MF dichotomy, none can be fully reduced to it. Contrasting some of these dichotomies highlights the fact that the MB–MF dichotomy is not simply a quantitative formalism of those more qualitative theories but is, indeed, theoretically distinct from most (such as the hot–cold emotional dimension) and offers patchy coverage of others (such as the habitual–goal-directed framework).

What is lost

Considering other dichotomous frameworks highlights the multifaceted nature of learning and decision-making by highlighting the many independent axes along which behaviour can be described. Although aligning cognitive, neural and behavioural data across various dualities offers the means to expose and examine key variables, something is necessarily lost when a system as complex as the brain is scrutinized through a dichotomous lens. Indeed, broad dichotomous descriptions (such as System1/System2) often lack predictive precision, and a proliferation of isolated contrastive frameworks stands to impair a coherent understanding of brain and behaviour[106–108]. The application of RL in studying the brain has facilitated notable progress by offering a formal framework through which theorems may be proved[109], axiomatic patterns may be described[110], brain function can be probed[29] and theories may be falsified. However, distilling learning and decision-making into a single MB–MF dimension risks conflating many other sources of variance, and, more importantly, threatens to dilute the formal merits of the computational RL framework to that of a verbal theory (for example, the agent 'uses' a 'model').

Paths forward

Identifying the computational primitives that support learning is an essential question at the core of cognitive (neuro) science, but also has implications for all domains that rely on learning — including education, public health, software design and so on. Characterizing these primitives will be an important step towards gaining deeper insight into learning differences between and across populations, including differences with developmental trajectories[111], differences depending on environmental factors and differences associated with 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 endeavour of deconstructing learning into its interpretable, neurally implementable basic primitives.

Disparities or inconsistencies between classic psychological theoretical frameworks offer opportunities to refine our understanding of their underlying computational primitives. For example, the apparent gaps between MB–MF RL and goal-directed–habitual behaviour could promote both theoretical and experimental advances. Failure to elicit a detectable change in the post-devaluation response rate using a devaluation protocol (that is, habits) could be caused by various investigable mechanisms (including degradation of the transition model, compromised goal maintenance or engagement of a MF controller). These gaps point to the importance of considering other dimensions of learning and decision-making (such as the potential role of value-free response[113]), and other facets of behaviour such as exploration that may be interpreted as being more MB or MF[14,56].

Computer science research (see FIG. 1) can aid us in the identification of additional relevant dimensions of learning. For example, algorithms have used hierarchical organization as a means of embedding task abstraction, whereby the agent can focus its decision-making machinery on high-level actions, such as 'make coffee', without needing to explicitly consider all of the subordinate steps involved. In hierarchical RL, information is learned and decisions are made at multiple levels of abstraction in parallel. Thus, hierarchical RL offers potentially beneficial task abstractions that can span across time[114–116] or the state or action space, and has been observed in humans[54,63,74,117,118]. Notably, hierarchical RL may be implemented using either MB planning or MF responses, which offers a rich set of computational tools that research (and the brain) may draw upon, 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, whereby 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 behaviour is shaped by, and exhibits interactions between, instructed and experienced trajectories through an environment[39]. Proposals have outlined frameworks in which supervised, unsupervised and RL systems interact to build and act on representations of the environment[119,120], further bending the notion that a singular spectrum of MB–MF control can sufficiently explain behaviour. A third algorithmic dimension that warrants consideration, as it may compound worries of misattribution, is the distinction between offline and online learning. Online-learning agents integrate observations as they arrive, whereas offline learners can use information at a later point for 'batch' updating, relying heavily on information storage and the ability to draw from it[23]. Offline learning has been suggested to occur between learning trials that involve working memory or hippocampal replay[121,122], or during consolidation in sleep[123], and may contribute to both model and reward learning (for example, the Dyna learning algorithm in which past experiences are virtually replayed to accelerate learning[23]).

Insights garnered from neuroscience should also continue contributing to enrich our understanding of the dimensions of learning and decision-making, as studies of regional specificity have implicated separable aspects of behaviour in different brain 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 to expected underlying neural systems[88,125,126]. Further examples of using insights from neuroscience to illuminate the computations underlying learning behaviour 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 work has begun to probe how these two systems may compete for control[91] or collaborate[130]. Such collaboration may emerge through relational associations maintained in the hippocampus upon which value may be learned[131,132], or through developing a representation that captures a transition structure in the environment[133]. Further strengthening a functional relationship between MB and MF RL, research has also offered evidence of a cooperative computational role between systems during reward learning as a means of actively sampling previous events to improve value estimates[93–95].

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, these components often interact in complex ways beyond competition for choice[134]. For example, in the MB–MF framework[4], striatal signals show that MB information is not segregated from MF reward prediction error. Similar findings have also been observed in recordings from

DA neurons[135,136]. Even functions known to stem from largely separable neural systems exhibit such interactions; for example, information in working memory seems to influence the reward prediction error computations of MF RL[89,124–126]. Going beyond simple dichotomies will necessitate not only increasing the dimensionality of the space of learning processes we consider but also considering how different dimensions interact.

In summary, there are numerous axes along which learning and decision-making vary as identified through various traditions of research (including psychology, AI and neuroscience). Future research should continue the progress of recent work in identifying many additional dimensions of learning that 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 versus habitual or otherwise.

Conclusions

We have attempted to show the importance of identifying the primitive components that support learning and decision-making as well as the risks inherent to compressing complex and multifaceted processes into a two-dimensional space. Although dual-system theories offer a means through which unique and dissociable components of learning and decision-making may be highlighted, key aspects could be fundamentally misattributed to unrelated computations, and scientific debate could become counterproductive when different subfields use the same label, even those as well computationally defined 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 towards a range of terms synonymous to many individuals with various dichotomous approximations that may or may not touch on shared functional or neural mechanisms. We have argued that such a shift leads to a dangerous approximation of a much higher-dimensional space. The rigour of computationally defined theories should not hide their limitations: the equations of a computational model are defined in a precise environment and do not necessarily expand seamlessly to capture neighbouring 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 concerned with applying learning science to other domains, such as AI or education. However, for research that aims to understand matters that are dependent on the mechanisms of learning (that is, the brain's implementation), such as the study of individual differences in learning, it is particularly important to ensure that the high-level theory of learning primitives proposes computational primitives that precisely relate to the underlying circuits. This effort may benefit from a renewed enthusiasm from computational modellers for the basic building blocks of psychology and neuroscience[143,144], and a better appreciation for the functional building blocks formalized by a rich computational theory.

Anne G. E. Collins[1]✉ *and Jeffrey Cockburn*[2]

[1]*Department of Psychology and the Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, CA, USA.*

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

✉*e-mail: annecollins@berkeley.edu*

https://doi.org/10.1038/s41583-020-0355-6

Published online 1 September 2020

1. Roiser, J. P. & Sahakian, B. J. Hot and cold cognition in depression. *CNS Spectr.* **18**, 139–149 (2013).
2. Dickinson, A. Actions and habits: the development of behavioural autonomy. *Philos. Trans. R. Soc. London. B Biol. Sci.* **308**, 67–78 (1985).
3. Sloman, S. A. The empirical case for two systems of reasoning. *Psychol. Bull.* **119**, 3 (1996).
4. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
5. Stanovich, K. E. & West, R. F. Individual differences in reasoning: implications for the rationality debate? *Behav. Brain Sci.* **23**, 645–665 (2000).
6. Kahneman, D. & Frederick, S. in *Heuristics and Biases: The Psychology of Intuitive Judgment* Ch. 2 (eds Gilovich, T., Griffin, D. & Kahneman, D.) 49–81 (Cambridge Univ. Press, 2002).
7. Daw, N. in *Decision Making, Affect, and Learning: Attention and Performance XXIII* Ch. 1 (eds Delgado, M. R., Phelps, E. A. and Robbins, T. W.) 1–26 (Oxford Univ. Press, 2011).
8. Marr, D. & Poggio, T. A computational theory of human stereo vision. *Proc. R. Soc. Lond. B. Biol. Sci.* **204**, 301–328 (1979).
9. Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D. & Daw, N. D. Model-based choices involve prospective neural activity. *Nat. Neurosci.* **18**, 767–772 (2015).
10. Daw, N. D. Are we of two minds? *Nat. Neurosci.* **21**, 1497–1499 (2018).
11. Dayan, P. Goal-directed control and its antipodes. *Neural Netw.* **22**, 213–219 (2009).
12. da Silva, C. F. & Hare, T. A. 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).
13. Moran, R., Keramati, M., Dayan, P. & Dolan, R. J. Retrospective model-based inference guides model-free credit assignment. *Nat. Commun.* **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 Comput. Biol.* **11**, e1004648 (2015).
15. Shahar, N. et al. Credit assignment to state-independent task representations and its relationship with model-based decision making. *Proc. Natl Acad. Sci. USA* **116**, 15871–15876 (2019).
16. Deserno, L. & Hauser, T. U. Beyond a cognitive dichotomy: can multiple decision systems prove useful to distinguish compulsive and impulsive symptom dimensions? *Biol. Psychiatry* https://doi.org/10.1016/j.biopsych.2020.03.004 (2020).
17. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
18. Dabney, W. et al. A distributional code for value in dopamine-based reinforcement learning. *Nature* **577**, 671–675 (2020).
19. Thorndike, E. L. *Animal Intelligence: Experimental Studies* (Transaction, 1965).
20. Bush, R. R. & Mosteller, F. *Stochastic models for learning* (John Wiley & Sons, Inc. 1955).
21. Pearce, J. M. & Hall, G. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* **87**, 532–552 (1980).
22. Rescorla, R. A. & Wagner, A. R. in *Classical Conditioning II: Current Research and Theory* Ch. 3 (eds Black, A. H. & Prokasy, W. F) 64–99 (Appleton-Century-Crofts, 1972).
23. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An Introduction* (MIT Press, 2018).
24. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947 (1996).
25. Bayer, H. M. & Glimcher, P. W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141 (2005).
26. Morris, G., Nevet, A., Arkadir, D., Vaadia, E. & Bergman, H. Midbrain dopamine neurons encode decisions for future action. *Nat. Neurosci.* **9**, 1057–1063 (2006).
27. Roesch, M. R., Calu, D. J. & Schoenbaum, G. Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat. Neurosci.* **10**, 1615–1624 (2007).
28. Shen, W., Flajolet, M., Greengard, P. & Surmeier, D. J. Dichotomous dopaminergic control of striatal synaptic plasticity. *Science* **321**, 848–851 (2008).
29. Steinberg, E. E. et al. A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* **16**, 966–973 (2013).
30. Kim, K. M. et al. Optogenetic mimicry of the transient activation of dopamine neurons by natural reward is sufficient for operant reinforcement. *PLoS ONE* **7**, e33612 (2012).
31. O'Doherty, J. et al. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004).
32. McClure, S. M., Berns, G. S. & Montague, P. R. Temporal prediction errors in a passive learning task activate human striatum. *Neuron* **38**, 339–346 (2003).
33. Samejima, K., Ueda, Y., Doya, K. & Kimura, M. Representation of action-specific reward values in the striatum. *Science* **310**, 1337–1340 (2005).
34. Lau, B. & Glimcher, P. W. Value representations in the primate striatum during matching behavior. *Neuron* **58**, 451–463 (2008).
35. Frank, M. J., Seeberger, L. C. & O'Reilly, R. C. By carrot or by stick: cognitive reinforcement learning in Parkinsonism. *Science* **306**, 1940–1943 (2004).
36. Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T. & Hutchison, K. E. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proc. Natl Acad. Sci. USA* **104**, 16311–16316 (2007).

37. Cockburn, J., Collins, A. G. & Frank, M. J. A reinforcement learning mechanism responsible for the valuation of free choice. *Neuron* **83**, 551–557 (2014).
38. Frank, M. J., O'Reilly, R. C. & Curran, T. When memory fails, intuition reigns: midazolam enhances implicit inference in humans. *Psychol. Sci.* **17**, 700–707 (2006).
39. Doll, B. B., Hutchison, K. E. & Frank, M. J. Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *J. Neurosci.* **31**, 6188–6198 (2011).
40. Doll, B. B. et al. Reduced susceptibility to confirmation bias in schizophrenia. *Cogn. Affect. Behav. Neurosci.* **14**, 715–728 (2014).
41. Berridge, K. C. The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology* **191**, 391–431 (2007).
42. Hamid, A. A. et al. Mesolimbic dopamine signals the value of work. *Nat. Neurosci.* **19**, 117–126 (2016).
43. Sharpe, M. J. et al. Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nat. Neurosci.* **20**, 735–742 (2017).
44. Tolman, E. C. Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208 (1948).
45. Economides, M., Kurth-Nelson, Z., Lübbert, A., Guitart-Masip, M. & Dolan, R. J. Model-based reasoning in humans becomes automatic with training. *PLoS Comput. Biol.* **11**, e1004463 (2015).
46. Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A. & Daw, N. D. Working-memory capacity protects model-based learning from stress. *Proc. Natl Acad. Sci. USA* **110**, 20941–20946 (2013).
47. Wunderlich, K., Smittenaar, P. & Dolan, R. J. Dopamine enhances model-based over model-free choice behavior. *Neuron* **75**, 418–424 (2012).
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. USA* **112**, 1595–1600 (2015).
49. Gillan, C. M., Otto, A. R., Phelps, E. A. & Daw, N. D. Model-based learning protects against forming habits. *Cogn. Affect. Behav. Neurosci.* **15**, 523–536 (2015).
50. Groman, S. M., Massi, B., Mathias, S. R., Lee, D. & Taylor, J. R. Model-free and model-based influences in addiction-related behaviors. *Biol. Psychiatry* **85**, 936–945 (2019).
51. Doll, B. B., Simon, D. A. & Daw, N. D. The ubiquity of model-based reinforcement learning. *Curr. Opin. Neurobiol.* **22**, 1075–1081 (2012).
52. Cushman, F. & Morris, A. Habitual control of goal selection in humans. *Proc. Natl Acad. Sci. USA* **112**, 201506367 (2015).
53. O'Reilly, R. C. & Frank, M. J. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput.* **18**, 283–328 (2006).
54. Collins, A. G. & Frank, M. J. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychol. Rev.* **120**, 190–229 (2013).
55. Momennejad, I. et al. The successor representation in human reinforcement learning. *Nat. Hum. Behav.* **1**, 680–692 (2017).
56. Da Silva, C. F. & Hare, T. A. Humans are primarily model-based and not model-free learners in the two-stage task. *bioRxiv* https://doi.org/10.1101/682922 (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, M. S., Murakami, M., Mainen, Z. F. & Dayan, P. An effect of serotonergic stimulation on learning rates for rewards apparent after long intertrial intervals. *Nat. Commun.* **9**, 2477 (2018).
59. Mohr, H. et al. Deterministic response strategies in a trial-and-error learning task. *PLoS Comput. Biol.* **14**, e1006621 (2018).
60. Hampton, A. N., Bossaerts, P. & O'Doherty, J. P. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* **26**, 8360–8367 (2006).
61. Boorman, E. D., Behrens, T. E. & Rushworth, M. F. Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. *PLoS Biol.* **9**, e1001093 (2011).
62. Behrens, T. E., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).
63. Collins, A. G. E. & Koechlin, E. Reasoning, learning, and creativity: frontal lobe function and human decision-making. *PLoS Biol.* **10**, e1001293 (2012).
64. Gershman, S. J., Norman, K. A. & Niv, Y. Discovering latent causes in reinforcement learning. *Curr. Opin. Behav. Sci.* **5**, 43–50 (2015).
65. Badre, D., Kayser, A. S. & Esposito, M. D. Article frontal cortex and the discovery of abstract action rules. *Neuron* **66**, 315–326 (2010).
66. Konovalov, A. & Krajbich, I. Mouse tracking reveals structure knowledge in the absence of model-based choice. *Nat. Commun.* **11**, 1893 (2020).
67. Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).
68. Huys, Q. J. et al. Interplay of approximate planning strategies. *Proc. Natl Acad. Sci. USA* **112**, 3098–3103 (2015).
69. Suzuki, S., Cross, L. & O'Doherty, J. P. Elucidating the underlying components of food valuation in the human orbitofrontal cortex. *Nat. Neurosci.* **20**, 1786 (2017).
70. Badre, D., Doll, B. B., Long, N. M. & Frank, M. J. Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron* **73**, 595–607 (2012).
71. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore–exploit dilemma. *J. Exp. Psychol. Gen.* **143**, 2074 (2014).
72. Otto, A. R., Gershman, S. J., Markman, A. B. & Daw, N. D. The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychol. Sci.* **24**, 751–761 (2013).
73. Niv, Y. et al. Reinforcement learning in multidimensional environments relies on attention mechanisms. *J. Neurosci.* **35**, 8145–8157 (2015).
74. Badre, D. & Frank, M. J. Mechanisms of hierarchical reinforcement learning in cortico-striatal circuits 2: evidence from fMRI. *Cereb. Cortex* **22**, 527–536 (2012).
75. Collins, A. G. E. Reinforcement learning: bringing together computation and cognition. *Curr. Opin. Behav. Sci.* **29**, 63–68 (2019).
76. Collins, A. G. in *Goal-directed Decision Making* (eds Morris, R., Bornstein, A. & Shenhav, A) 105–123 (Elsevier, 2018).
77. Donoso, M., Collins, A. G. E. & Koechlin, E. Foundations of human reasoning in the prefrontal cortex. *Science* **344**, 1481–1486 (2014).
78. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–278 (2014).
79. Schuck, N. W., Wilson, R. & Niv, Y. in *Goal-directed Decision Making* (eds Morris, R., Bornstein, A. & Shenhav, A) 259–278 (Elsevier, 2018).
80. Ballard, I. C., Wagner, A. D. & McClure, S. M. Hippocampal pattern separation supports reinforcement learning. *Nat. Commun.* **10**, 1073 (2019).
81. Redish, A. D., Jensen, S., Johnson, A. & Kurth-Nelson, Z. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychol. Rev.* **114**, 784 (2007).
82. Bouton, M. E. Context and behavioral processes in extinction. *Learn. Mem.* **11**, 485–494 (2004).
83. Rescorla, R. A. Spontaneous recovery. *Learn. Mem.* **11**, 501–509 (2004).
84. O'Reilly, R. C., Frank, M. J., Hazy, T. E. & Watz, B. PVLV: the primary value and learned value Pavlovian learning algorithm. *Behav. Neurosci.* **121**, 31 (2007).
85. Gershman, S. J., Blei, D. M. & Niv, Y. Context, learning, and extinction. *Psychol. Rev.* **117**, 197–209 (2010).
86. Wang, J. X. et al. Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* **21**, 860–868 (2018).
87. Iigaya, K. et al. Deviation from the matching law reflects an optimal strategy involving learning over multiple timescales. *Nat. Commun.* **10**, 1466 (2019).
88. Collins, A. G. E. & Frank, M. J. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *Eur. J. Neurosci.* **35**, 1024–1035 (2012).
89. Collins, A. G. E. The tortoise and the hare: interactions between reinforcement learning and working memory. *J. Cogn. Neurosci.* **30**, 1422–1432 (2017).
90. Viejo, G., Girard, B. B., 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 (2017).
91. Poldrack, R. A. et al. Interactive memory systems in the human brain. *Nature* **414**, 546–550 (2001).
92. Foerde, K. & Shohamy, D. Feedback timing modulates brain systems for learning in humans. *J. Neurosci.* **31**, 13157–13167 (2011).
93. Bornstein, A. M., Khaw, M. W., Shohamy, D. & Daw, N. D. Reminders of past choices bias decisions for reward in humans. *Nat. Commun.* **8**, 15958 (2017).
94. Bornstein, A. M. & Norman, K. A. Reinstated episodic context guides sampling-based decisions for reward. *Nat. Neurosci.* **20**, 997–1003 (2017).
95. Vikbladh, O. M. et al. Hippocampal contributions to model-based planning and spatial memory. *Neuron* **102**, 683–693 (2019).
96. Decker, J. H., Otto, A. R., Daw, N. D. & Hartley, C. A. From creatures of habit to goal-directed learners: tracking the developmental emergence of model-based reinforcement learning. *Psychol. Sci.* **27**, 848–858 (2016).
97. Dickinson, A. & Balleine, B. Motivational control of goal-directed action. *Anim. Learn. Behav.* **22**, 1–18 (1994).
98. Balleine, B. W. & Dickinson, A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* **37**, 407–419 (1998).
99. Daw, N. D. & Doya, K. The computational neurobiology of learning and reward. *Curr. Opin. Neurobiol.* **16**, 199–204 (2006).
100. Friedel, E. et al. Devaluation and sequential decisions: linking goal-directed and model-based behavior. *Front. Hum. Neurosci.* **8**, 587 (2014).
101. de Wit, S. et al. Shifting the balance between goals and habits: five failures in experimental habit induction. *J. Exp. Psychol. Gen.* **147**, 1043–1065 (2018).
102. Madrigal, R. Hot vs. cold cognitions and consumers' reactions to sporting event outcomes. *J. Consum. Psychol.* **18**, 304–319 (2008).
103. Peterson, E. & Welsh, M. C. in *Handbook of Executive Functioning* (eds Goldstein, S. & Naglieri, J. A.) 45–65 (Springer, 2014).
104. Barch, D. M. et al. Explicit and implicit reinforcement learning across the psychosis spectrum. *J. Abnorm. Psychol.* **126**, 694–711 (2017).
105. Taylor, J. A., Krakauer, J. W. & Ivry, R. B. Explicit and implicit contributions to learning in a sensorimotor adaptation task. *J. Neurosci.* **34**, 3023–3032 (2014).
106. Sloman, S. A. in *Heuristics and biases: The psychology of intuitive judgment* Ch. 22 (eds Gilovich, T., Griffin, D. & Kahneman D.) 379–396 (Cambridge Univ. Press, 2002).
107. Evans, J. S. B. T. in In two minds: Dual processes and beyond (eds J. S. B. T. Evans & K. Frankish) p. 33–54 (Oxford Univ. Press, 2009).
108. Stanovich, K. *Rationality and the Reflective Mind* (Oxford Univ. Press, 2011).
109. Dayan, P. The convergence of TD(λ) for general λ. *Mach. Learn.* **8**, 341–362 (1992).
110. Caplin, A. & Dean, M. Axiomatic methods, dopamine and reward prediction error. *Curr. Opin. Neurobiol.* **18**, 197–202 (2008).
111. van den Bos, W., Bruckner, R., Nassar, M. R., Mata, R. & Eppinger, B. Computational neuroscience across the lifespan: promises and pitfalls. *Dev. Cogn. Neurosci.* **33**, 42–53 (2018).
112. Adams, R. A., Huys, Q. J. & Roiser, J. P. Computational psychiatry: towards a mathematically informed understanding of mental illness. *J. Neurol. Neurosurg. Psychiatry* **87**, 53–63 (2016).
113. Miller, K. J., Shenhav, A. & Ludvig, E. A. Habits without values. *Psychol. Rev.* **126**, 292–311 (2019).
114. Botvinick, M. M., Niv, Y. & Barto, A. Hierarchically organized behavior and its neural foundations: a reinforcement-learning perspective. *Cognition* **113**, 262–280 (2009).
115. Konidaris, G. & Barto, A. G. in *Advances in Neural Information Processing Systems 22* (eds Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I. & Culotta, A.) 1015–1023 (NIPS, 2009).
116. Konidaris, G. On the necessity of abstraction. *Curr. Opin. Behav. Sci.* **29**, 1–7 (2019).
117. Frank, M. J. & Fossella, J. A. Neurogenetics and pharmacology of learning, motivation, and cognition. *Neuropsychopharmacology* **36**, 133–152 (2010).
118. Collins, A. G. E., Cavanagh, J. F. & Frank, M. J. Human EEG uncovers latent generalizable rule structure during learning. *J. Neurosci.* **34**, 4677–4685 (2014).

119. Doya, K. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw.* **12**, 961–974 (1999).
120. Fermin, A. S. et al. Model-based action planning involves cortico-cerebellar and basal ganglia networks. *Sci. Rep.* **6**, 31378 (2016).
121. Gershman, S. J., Markman, A. B. & Otto, A. R. Retrospective revaluation in sequential decision making: a tale of two systems. *J. Exp. Psychol. Gen.* **143**, 182 (2014).
122. Pfeiffer, B. E. & Foster, D. J. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* **497**, 74–79 (2013).
123. Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S. I. & Battaglia, F. P. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nat. Neurosci.* **12**, 919–926 (2009).
124. Collins, A. G. E., Albrecht, M. A., Waltz, J. A., Gold, J. M. & Frank, M. J. Interactions among working memory, reinforcement learning, and effort in value-based choice: a new paradigm and selective deficits in schizophrenia. *Biol. Psychiatry* **82**, 431–439 (2017).
125. Collins, A. G. E., Ciullo, B., Frank, M. J. & Badre, D. Working memory load strengthens reward prediction errors. *J. Neurosci.* **37**, 2700–2716 (2017).
126. Collins, A. A. G. E. & Frank, M. J. M. Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proc. Natl Acad. Sci. USA* **115**, 2502–2507 (2018).
127. Knowlton, B. J., Mangels, J. A. & Squire, L. R. A neostriatal habit learning system in humans. *Science* **273**, 1399–1402 (1996).
128. Squire, L. R. & Zola, S. M. Structure and function of declarative and nondeclarative memory systems. *Proc. Natl Acad. Sci. USA* **93**, 13515–13522 (1996).
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. Learn. Mem.* **96**, 624–636 (2011).
131. Wimmer, G. E., Daw, N. D. & Shohamy, D. Generalization of value in reinforcement learning by humans. *Eur. J. Neurosci.* **35**, 1092–1104 (2012).
132. Wimmer, G. E., Braun, E. K., Daw, N. D. & Shohamy, D. Episodic memory encoding interferes with reward learning and decreases striatal prediction errors. *J. Neurosci.* **34**, 14901–14912 (2014).
133. Gershman, S. J. The successor representation: its computational logic and neural substrates. *J. Neurosci.* **38**, 7193–7200 (2018).
134. Kool, W., Cushman, F. A. & Gershman, S. J. in *Goal-directed Decision Making* Ch. 7 (eds Morris, R. W. & Bornstein, A.) 153–178 (Elsevier, 2018).
135. Langdon, A. J., Sharpe, M. J., Schoenbaum, G. & Niv, Y. Model-based predictions for dopamine. *Curr. Opin. Neurobiol.* **49**, 1–7 (2018).
136. Starkweather, C. K., Babayan, B. M., Uchida, N. & Gershman, S. J. Dopamine reward prediction errors reflect hidden-state inference across time. *Nat. Neurosci.* **20**, 581–589 (2017).
137. Krueger, K. A. & Dayan, P. Flexible shaping: how learning in small steps helps. *Cognition* **110**, 380–394 (2009).
138. Bhandari, A. & Badre, D. Learning and transfer of working memory gating policies. *Cognition* **172**, 89–100 (2018).
139. Leong, Y. C. et al. Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron* **93**, 451–463 (2017).
140. Farashahi, S., Rowe, K., Aslami, Z., Lee, D. & Soltani, A. Feature-based learning improves adaptability without compromising precision. *Nat. Commun.* **8**, 1768 (2017).
141. Bach, D. R. & Dolan, R. J. Knowing how much you don't know: a neural organization of uncertainty estimates. *Nat. Rev. Neurosci.* **13**, 572–586 (2012).
142. Pulcu, E. & Browning, M. The misestimation of uncertainty in affective disorders. *Trends Cogn. Sci.* **23**, 865–875 (2019).
143. Badre, D., Frank, M. J. & Moore, C. I. Interactionist neuroscience. *Neuron* **88**, 855–860 (2015).
144. Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D. Neuroscience needs behavior: correcting a reductionist bias. *Neuron* **93**, 480–490 (2017).
145. Doll, B. B., Shohamy, D. & Daw, N. D. Multiple memory systems as substrates for multiple decision systems. *Neurobiol. Learn. Mem.* **117**, 4–13 (2014).
146. Smittenaar, P., FitzGerald, T. H., Romei, V., Wright, N. D. & Dolan, R. J. Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron* **80**, 914–919 (2013).
147. Doll, B. B., Bath, K. G., Daw, N. D. & Frank, M. J. Variability in dopamine genes dissociates model-based and model-free reinforcement learning. *J. Neurosci.* **36**, 1211–1222 (2016).
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).
149. Voon, V., Reiter, A., Sebold, M. & Groman, S. Model-based control in dimensional psychiatry. *Biol. Psychiatry* **82**, 391–400 (2017).
150. Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A. & Daw, N. D. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife* **5**, e11305 (2016).
151. Culbreth, A. J., Westbrook, A., Daw, N. D., Botvinick, M. & Barch, D. M. Reduced model-based decision-making in schizophrenia. *J. Abnorm. Psychol.* **125**, 777–787 (2016).
152. Patzelt, E. H., Kool, W., Millner, A. J. & Gershman, S. J. Incentives boost model-based control across a range of severity on several psychiatric constructs. *Biol. Psychiatry* **85**, 425–433 (2019).
153. Skinner, B. F. *The Selection of Behavior: The Operant Behaviorism of BF Skinner: Comments and Consequences* (CUP Archive, 1988).
154. Corbit, L. H., Muir, J. L. & Balleine, B. W. Lesions of mediodorsal thalamus and anterior thalamic nuclei produce dissociable effects on instrumental conditioning in rats. *Eur. J. Neurosci.* **18**, 1286–1294 (2003).
155. Coutureau, E. & Killcross, S. Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. *Behav. Brain Res.* **146**, 167–174 (2003).
156. Yin, H. H., Knowlton, B. J. & Balleine, B. W. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur. J. Neurosci.* **19**, 181–189 (2004).
157. Yin, H. H., Knowlton, B. J. & Balleine, B. W. Inactivation of dorsolateral striatum enhances sensitivity to changes in the action–outcome contingency in instrumental conditioning. *Behav. Brain Res.* **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).

Author contributions
The authors contributed equally to all aspects of the article.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Neuroscience thanks the anonymous reviewer(s) for their contribution to the peer review of this work.