Competition and Cooperation Between Multiple Reinforcement Learning Systems

Wouter Kool, Fiery A. Cushman, & Samuel J. Gershman

Most psychological research on reinforcement learning has depicted two systems locked in battle for control of behavior: a flexible but computationally expensive "model-based" system and an inflexible but cheap "model-free" system. However, the complete picture is more complex, with the two systems cooperating in myriad ways. We focus on two issues at the frontier of this research program. First, how is the conflict between these systems adjudicated? Second, how can the systems be combined to harness the relative strengths of each? This chapter reviews recent work on competition and cooperation between the two systems, highlighting the computational principles that govern different forms of interaction.

As you leave work each day, how do you choose a route home? Prominent dual-system accounts posit two distinct cognitive systems that solve this task in different ways (Balleine & O'Doherty, 2009; Dickinson, 1985; Fudenberg & Levine, 2006; Kahneman, 1973; Sloman, 1996). On the one hand, you could decide your route home by relying on *habit*. Since you have successfully taken one particular route to your house many times, this route has been ingrained into your motor system, and can be executed quickly and automatically. Habits are useful because they make often-repeated behavior efficient and automatized; however, they are also inflexible and therefore more likely to produce errors. For example, consider the case where your significant other asked you to buy some toilet paper on your way back home. In this case, it would be better to suppress the habitual route and engage in *goal-directed* control. This involves the recall of the alternate goal (picking up toilet paper), and planning a new route that goes past the convenience store, using an internal model ("cognitive map") of the environment. Goal-directed planning is useful because it is more flexible and consequently more accurate than relying on habit. However, it also carries significant computational costs (Gershman & Daw, 2012).

These two systems are typically theorized as *competitors*, vying for control of behavior. A major goal of modern decision research is understanding how control is allocated between the two systems. We will attempt to summarize and extend this line of research.

Yet, the two systems may also interact *cooperatively*. For example, you might learn a habit to check traffic reports before you leave work, because this facilitates planning an optimal route. Moreover, the act of "checking" could involve elements of goal-directed planning—for instance, searching for radio stations—even if initiated out of habit. These illustrate just two forms of cooperation: habitual actions can support effective goal pursuit, and even drive the selection of goals themselves.

Until recently, the computational principles underlying the competition and cooperation between habitual and goal-directed systems were poorly understand. Armed with a new set of sequential decision tasks, researchers are now able to track habitual and goal-directed influences on behavior across an experimental session (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Doll, Duncan, Simon, Shohamy, & Daw, 2015; Keramati, Smittenaar, Dolan, & Dayan, 2016; Kool, Cushman, & Gershman, 2016). This work has spurred new computational approaches to multi-system reinforcement learning and control architectures.

In this chapter, we review recent work on both competition and cooperation. First, we will provide a short, non-technical exposition of the computational framework underlying this research (see Gershman, 2017 for a technical review). Next, we will discuss recent work that suggests how competition between habit and planning can be understood as a cost-benefit tradeoff. Finally, we describe several studies that detail how the complementary strengths of habitual and goal-directed systems can be combined cooperatively to achieve both efficiency and accuracy.

Model-free and model-based control in reinforcement learning

The core problem in reinforcement learning is estimating the *value* (expected discounted return) of state-action pairs in order to guide action selection. Broadly speaking, there are two strategies for solving this problem: a model-free strategy that estimates values incrementally from experience, and a model-based strategy that learns a world model (reward and transition functions) which can then be used to plan an optimal policy. A central tenet of modern reinforcement learning theory posits that the model-free strategy is implemented by the habitual system, and the model-based strategy is implemented by the goal-directed system (Daw, Niv, & Dayan, 2005; Dolan & Dayan, 2013).

Roughly speaking, the model-free strategy is a form of Thorndike's Law of Effect, which states that actions that led to a reward become more likely to be repeated (Thorndike, 1911). This strategy is referred to as "model-free" because it does not rely on an internal model of the environment. Instead, values are stored in a cached format (a look-up table or function approximator), which allows them to be quickly retrieved. These values can be updated incrementally using simple error-driven learning rules like the temporal difference learning algorithm (Sutton & Barto, 1998). The main downside of the model-free strategy is its inflexibility: when a change in the environment or task occurs, the entire set of cached values needs to be relearned through experience. This inflexibility, ingrained by repetition, is what makes the model-free strategy habitual. In summary, the model-free strategy achieves efficiency of learning and control at the expense of flexibility in the face of change.

The model-based strategy, by contrast, represents its knowledge in the form of an internal model that can be modified locally when changes occur (e.g., if a particular route is blocked, only that part of the model is modified). These local changes can then induce global effects on the value function, which is computed on the fly using planning or dynamic programming algorithms. Thus, the model-based strategy, unlike the model-free strategy, need not cache values. As a consequence, the model-based strategy can flexibly modify its policy in pursuit of a goal without relearning the entire model. This flexibility is only available at a computational cost, however, since model-based algorithms are

inevitably more time and resource intensive than querying a look-up table of cached values or function approximator (Daw et al., 2005; Keramati, Dezfouli, & Piray, 2011).

Principles of competition

Distinguishing habit from planning in humans

A long line of research in psychology and neuroscience has sought empirical evidence for the distinction between these model-free and model-based reinforcement learning systems. Early studies tended to focus on animal models, and this literature has been reviewed extensively elsewhere (Dolan & Dayan, 2013; Gershman, 2017) so we will not cover it here. Instead, we focus on more recent studies with human subjects. We will describe how one particular experimental paradigm, a sequential decision task which we will refer to as the "Daw two-step task" (Daw et al., 2011), has been pivotal in revealing the competition between model-free and model-based control in humans. We then turn to our main topic of interest in this section: Given that the systems can compete for control, how is this competition arbitrated?

Many recent studies of model-free and model-based control in humans have used the Daw two-step (Daw et al., 2011), summarized in Figure 1 (following Decker, Otto, Daw, & Hartley, 2016, the depicted version of this task features a space travel cover story to make it more engaging for participants). The key appeal of this task is that it can be used to quantitatively distinguish the influence of model-free and model-based control on choices (though see Akam, Costa, & Dayan, 2015). Each trial of the Daw two-step task starts with a choice between two stimuli (spaceships), which lead probabilistically to one

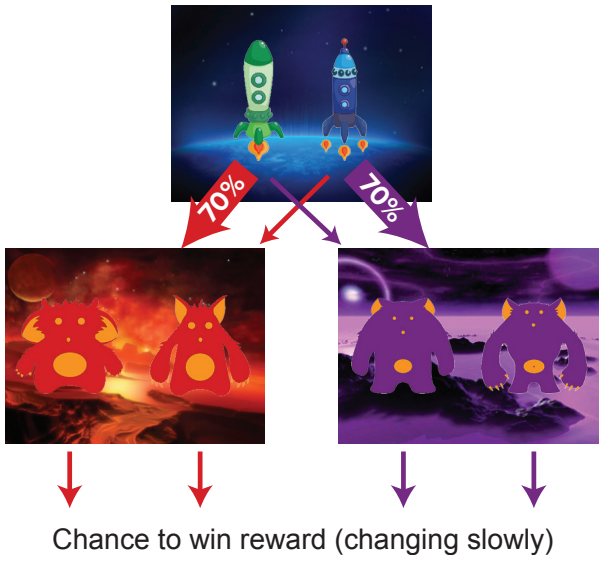

Figure 1. Design and state transition structure of Daw two-step task (Daw et al., 2011; Decker et al., 2016). Each first-stage choice has a high probability (70%) of transitioning to one of two second-stage states and a low probability of transitioning to the other. Each second-stage choice is associated with a probability of obtaining a binary reward (between 0.25 and 0.75) that slowly changes across the duration of the experiment according to a Gaussian random walk with $\sigma = 0.025$.

of two second-stage states (planets). At these second-stage states, the participant then makes a second choice between two stimuli (aliens) that both offer a chance of obtaining a monetary reward (space treasure). The reward probabilities for these second-stage stimuli change slowly and independently throughout the task in order to encourage continuous learning. The most important feature of the Daw two-step task is its transition structure from the first-stage stimuli to the second-stage states. Specifically, each first-stage option leads to one of the second-stage states with a high probability (a "common" transition), whereas on a minority of the trials they lead to the other state (a "rare" transition).

Through these low-probability transitions between actions and rewards, the Daw two-step task can behaviorally distinguish between model-free and model-based choice. Because the model-free strategy does not have access to the task structure, it will increase the probability of taking the previous action if it led to reward, regardless of whether this was obtained through a common or a rare transition. Therefore, choice dictated by a purely model-free agent looks like a main effect of reward, with increased probability of repeating the previous action after a reward, and with no effect of the previous transition (Figure 2A). The model-based strategy, on the other hand, computes the first-stage action values through planning, using the transition structure to compute the expected value at the second stage for either action. Therefore, this system will reduce the likelihood of repeating the first-stage action after a reward obtained through a rare transition, since the other first-stage action has a higher likelihood to lead to the previously rewarded second-stage state. This behavior is reflected as a cross-over interaction between the previous transition type and previous reward on the probability of staying: after rare transitions, wins predict a switch and losses predict a stay (Figure 2B).

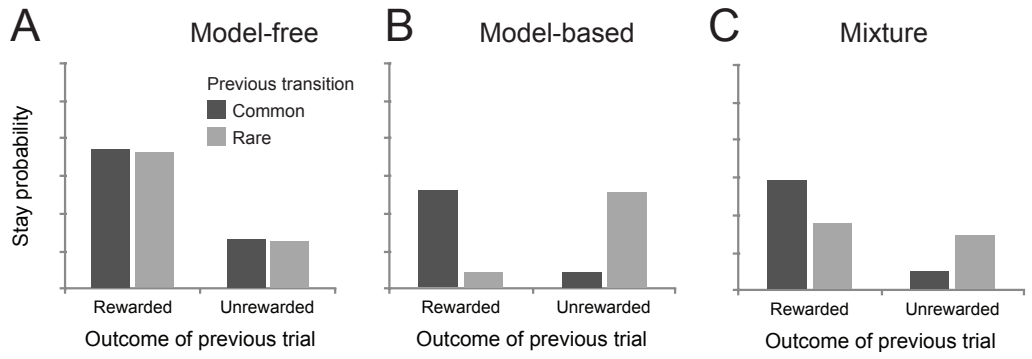

Figure 2. Reprinted from Kool et al. (2016). Probability of repeating the first-stage choice for three agents. (A) For model-free agents, the probability of repeating the previous choice is dependent only on whether a reward was obtained, and not on transition structure. (B) Model-based behavior is reflected in an interaction between previous transition and outcome, increasing the probability of transitioning to the state where reward was obtained. (C) Behavioral performance on this task reflects features of both model-based and model-free decision making, the main effect of previous reward and its interaction with the previous transition.

Interestingly, behavioral performance on the Daw two-step task reflects a mixture of these strategies (Figure 2C). The probability of repeating the previous actions shows both the

model-free main effect of previous reward but also the model-based crossover interaction between previous transition type and previous reward. The relative influence of the model-based and model-free systems on this task can be estimated by fitting a reinforcement-learning model to participants' behavior. Here, both strategies compute first-stage action values, which are then combined according to a weight parameter that determines the relative balance between model-free and model-based control.

The relative balance between model-based and model-free control indexed by this task has been linked to a broad range of other cognitive, neural and clinical phenomena. For example, Decker, Otto, Daw and Hartley (2016) showed that children show virtually no signs of model-based control, and that our ability for model-based planning develops through adolescence into adulthood (see Chapter 13 by Hartley in this volume). Gillan, Kosinski, Whelan, Phelps, and Daw (2016) have reported that the degree of model-based control in this task positively predicts psychiatric symptoms related to compulsive behavior (see Chapter 15 by Morris and Chapter 16 by de Wit in this volume), and others have shown that it also negatively predicts personality traits such as alcohol dependence (Sebold et al., 2014; see Chapter 17 by Corbit in this volume) and extraversion (Skatova, Chan, & Daw, 2015).

In addition to these findings that bolster the applicability of the two-step task to the broader field of psychology, it can also account for important phenomena in the reinforcement learning literature, such as the finding that overtraining of an action-reward association induces insensitivity to subsequent outcome devaluation (a hallmark feature of habitual control; Gillan, Otto, Phelps, & Daw, 2015).

Arbitration between habit and planning as a cost-benefit tradeoff

The finding that people show a balance between model-based and model-free control on the Daw two-step task raises the question of whether and how people decide, from moment to moment, which strategy to use. Although there are several theoretical proposals on this topic (Boureau, Sokol-Hessner, & Daw, 2015; Gershman, Horvitz, & Tenenbaum, 2015; Griffiths, Lieder, & Goodman, 2015; Keramati et al., 2011; Pezzulo, Rigoli, & Chersi, 2013), it has received surprisingly little empirical focus (but see Daw et al., 2005; Lee, Shimojo, & O'Doherty, 2014).

Several experimental manipulations have been discovered to alter the balance between model-free and model-based control, and these provide key clues about the form and function of arbitration between RL systems. As we review, many of these implicate some form of executive function or working memory in model-based control. In one such case (Otto, Gershman, Markman, & Daw, 2013), participants performed the Daw two-step task while they were sometimes required to perform a numerical Stroop task that taxed their working memory and therefore reduced the amount of available cognitive resources. At the start of those trials, participants kept two numbers of different value and physical size in working memory. After the reward outcome of the two-step trial was presented, participants were then prompted to indicate on what side of the screen the number with larger size or value had appeared. Interestingly, on trials with this "load" condition, subjects showed a strong reliance on the model-free strategy, and virtually no influence of a model-based strategy (Otto, Gershman, et al., 2013). This study suggests that the exertion of model-based control relies, at least in part, on executive functioning or

cognitive control. This set of cognitive processes, which are dependent on computations in the frontal cortex, allow us to reconfigure information processing in order to execute novel and effortful tasks (Miller & Cohen, 2001).

Another clue for the involvement of executive functioning in model-based planning comes from a study by Smittenaar, FitzGerald, Romei, Wright and Dolan (2013). In this experiment, participants performed the Daw two-step task while activity in their right dorsolateral prefrontal cortex (dlPFC), a region which is critical for the functioning of cognitive control, was sometimes disrupted using transcranial magnetic stimulation. Interestingly, performance on the task showed increased reliance on habitual control during those trials, indicating a crucial role for the dlPFC and executive functioning in model-based planning (see also Gläscher, Daw, Dayan, & O'Doherty, 2010; Lee et al., 2014).

Several other reports have yielded consistent evidence, in the form of robust correlations between individual differences in the degree of model-based control used in the Daw two-step task and measures of cognitive control ability. For example, Otto et al. (Otto, Skatova, Madlon-Kay, & Daw, 2015), showed that people with reduced performance in a response conflict task (such as the Stroop task; Stroop, 1935) also showed reduced employment of model-based control. In another study, participants with increased working memory capacity showed a reduced shift towards model-free control under stress (Otto, Raio, Chiang, Phelps, & Daw, 2013). In addition, Schad et al. (2014) showed that measures of general intelligence predicted reliance on model-based control. Their participants completed both the Daw two-step task and also the Trail Making Task (Army Individual Test Battery, 1944), in which participants use a pencil to connect numbers and letters, randomly distributed on a sheet of paper, in ascending order, while also alternating between numbers and letters (i.e., 1-A-2-B-3-C, etc.). Interestingly, individuals with increased processing speed on this task, indicating increased ability for cognitive control in the form of task switching, also showed a greater reliance on model-based control in the Daw two-step task.

We now address the question of whether, and how, the brain arbitrates between model-based and model-free control. One potential metacontrol strategy would simply be to always use the more accurate model-based system when the necessary cognitive resources are available, and only use the habitual system when they are occupied or otherwise inoperative. Note that, although this would lead to increased average accuracy, such a model does not describe how its resources should be allocated when they could be devoted to multiple tasks. In other words, this model does not predict how people allocate control resources when the available tasks together demand more resources than are available.

When aiming to describe such a tradeoff, it would be sensible for a model to be sensitive to the elevated computational costs that are associated with model-based control, since those cognitive resources could be applied to other rewarding tasks. Consequently, we propose that allocation of control is based on the costs and benefits associated with each system in a given task. In this case, model-based control would be deployed when it generates enough of a reward advantage over model-free control to offset its costs.

Consistent with this possibility, recent experimental evidence suggests that demands for cognitive control register as intrinsically costly (Kool, McGuire, Rosen, & Botvinick,

2010; Schouppe, Ridderinkhof, Verguts, & Notebaert, 2014; Westbrook, Kester, & Braver, 2013). For example, in the demand selection task (Kool et al., 2010), participants freely choose between task options that require different amounts of cognitive control, and subsequently show a strong bias towards the lines of action with the smallest control demands. The intrinsic cost account predicts, in addition, that this avoidance bias should be offset by incentives. Indeed, several studies provide evidence for this hypothesis by showing increased willingness to perform demanding tasks when appropriately rewarded (Westbrook et al., 2013), even if this commits them to increased time towards goal attainment (Kool et al., 2010).

Based on these, and other, findings (for a review see Botvinick & Braver, 2015), recent accounts of executive functioning propose that the exertion of cognitive control can best be understood as a form of cost-benefit decision making. For example, Shenhav, Botvinick, and Cohen (2013) have proposed that the brain computes an 'expected value of control' for each action—the expected rewarded discounted by the cost of associated control demands—and then chooses the action with highest value. Other researchers have proposed similar accounts (Gershman et al., 2015; Griffiths et al., 2015), whereby metacontrol between different systems is determined by the 'value of computation', the expected reward for a given action subtracted by the costs of computation and time.

An older, but related, account was developed by Payne, Bettman, and Johnson (1988; 1993), who proposed that humans are 'adaptive decision makers', choosing among strategies by balancing accuracy against cognitive effort. Finally, a recent model from Kurzban, Duckworth, Kable, & Meyers (2013) addresses the cost-benefit tradeoff from a slightly different angle. They argue that the cost of effort, and therefore the subsequent implementation of control for a certain action, is dependent on the opportunity costs of the alternatively available actions. This model predicts that subjective experiences of effort, and subsequent reductions in control, depend on the value of the next-best line of action. In summary, even though these proposals differ in terms of how costs influence decision making, they all center on the idea that the mobilization of control can best be understood as a form of cost-benefit tradeoff.

Below we sketch our own recent efforts to combine these insights from reinforcement learning theory in general—and the Daw two-step task, in particular—with the emerging view of cognitive control as value-based decision making. We then review several other related approaches in the contemporary literature.

Control-reward tradeoff in the two-step task

We propose that arbitration between model-based and model-free control is achieved by integrating the costs and benefits of each system. The rewards obtained by each system can be calculated by observing the average returns obtained by each control system, independently, and conditioned on the present task. Next, the brain uses these resulting 'controller values' to select actions that maximize future cumulative reward. In doing so, it imposes an intrinsic, subject "cost" on the model-based controller. This cost represents the foregone reward due to model-based control, for instance due to the potentially longer decision time and due to the foregone opportunity to deploy limited cognitive control resources on other, concurrent tasks.

A core prediction of this model is that manipulating the rewards available during a decision-making task should alter the balance between model-free and model-based control. A natural candidate task to test this prediction is the Daw two-step task. Indeed, the model-based strategy in this task has been described as "optimal" (e.g., Sebold et al., 2014). Thus, one would predict that the more money at stake on any given trial of the task, the more willing the participant should be to pay the intrinsic cost of cognitive control in order to obtain the benefits of accurate performance.

In practice, however, recent research on this task shows that increase reliance on the model-based system does not predict increased performance accuracy on the Daw two-step task (Akam et al., 2015; Kool et al., 2016). To show this, Kool et al. (2016) recorded the average reward rate of many reinforcement learning agents that varied across a range from pure model-free control to pure model-based control (see Figure 3A). These simulations showed no systematic relationship between reward rate and model-based control for the original Daw two-step task, or for several related variants of this task (Dezfouli & Balleine, 2013; Doll et al., 2015) across a wide range of reinforcement learning parameters. Consistent with this simulation result, they also found no correlation between model-based control and average reward in a subsequent experiment (Figure 3B). The absence of this relation is produced by the interaction of at least five factors, several of which appear to prevent the model-based system from obtaining sufficiently reliable reward estimates (Kool et al., 2016). In short, the effectiveness of the model-based strategy is weakened on the Daw two-step task, because the first-stage choices carry relatively decreased importance, and because this strategy does not have access to accurate representations of the second-stage reward outcomes. The fact that there is no control-reward tradeoff in the Daw two-step task makes it ill-suited to test the cost-benefit hypothesis of reinforcement learning arbitration, for example by testing the effect of increased "stakes" on controller selection.

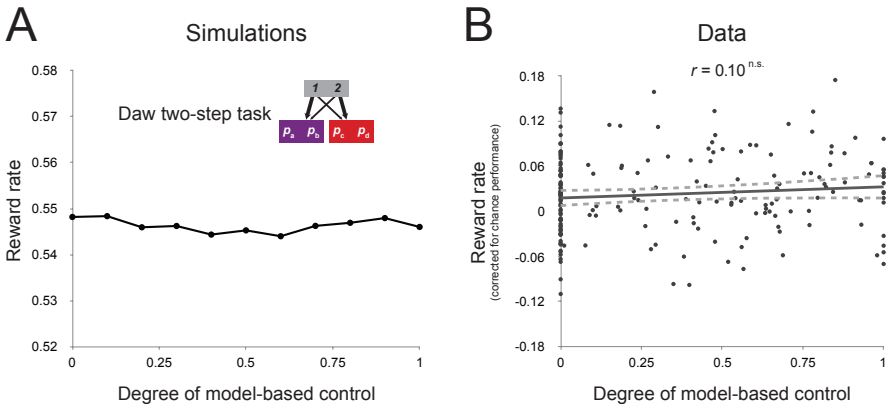

Figure 3. Adapted from Kool et al. (2016). Control-reward tradeoff in the Daw two-step task. (A) The relationship between the degree of model-based control and reward rate across 1000 simulations (with reinforcement learning parameters mirroring the median fits reported by Daw et al. (2011)). Importantly, these simulation results show that the task does not embody a trade-off between model-based control and reward. (B) Relationship between the estimated degree of model-based control and reward rate in the Daw two-step task (Daw et al., 2011). Consistent with simulation results, there is no correlation between these variables (n = 197). Dashed lines indicate the 95% confidence interval.

A novel two-step paradigm

In order to gain more experimental and computational traction on a control-reward tradeoff in reinforcement learning, Kool et al. (2016) developed a novel two-step task that theoretically and empirically achieves a tradeoff between control and reward. The changes in this new task are based on the factors that were identified to produce the absence of this relationship in the Daw two-step task. One of the more notable changes to this paradigm is that it adopts a different task structure (Figure 4; Doll et al., 2015). This task uses two first-stage states (randomly selected at the start of each trial) that both offer deterministic choices to one of two second-stage states. Both these second-stage states the choices again are associated with a reward outcome that randomly changes across the experimental session. Specifically, the drifting reward probabilities at the second stage are replaced with drifting scalar rewards (ranging from a negative to a positive number), so that the payoff of each action is identical to its value. This change was implemented to increase the informativeness of each reward outcome, and thus to increase model-based accuracy.

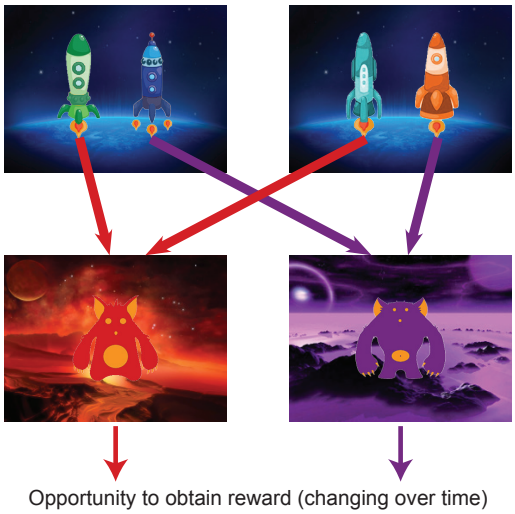

Figure 4. Design and state transition structure of the novel two-step task. Each first-stage choice deterministically transitions to one of two second-stage states. Each second-stage is associated with a scalar reward (between 0 and 9) which changes over the duration of the experiment according to a random Gaussian walk with $\sigma = 2$.

The dissociation between model-free and model-based control in this task follows a different logic than the Daw two-step task. Since the model-free system only learns state-action-reward outcomes, it will not be able to transfer information learned in one starting state to the other starting state. In other words, rewards that are obtained in one starting state only increase the likelihood of revisiting that second-stage when the next trial starts in the same starting state, but should not affect subsequent choices from the other starting state. The model-based system, on the other hand, treats the two starting states as functionally equivalent, because it realizes the implicit equivalence of their action

outcomes. Therefore, it will be able to generalize knowledge across them. So, reward outcomes at the second-stage should equally affect first-stage choices in the next trial, independent of whether this trial starts with the same state as the previous one.

This novel version of the two-step task incorporates many changes that increase the importance of the first-stage state and the ability of the system to learn the second-stage action values. Because of this, it achieves a tradeoff between control and reward. This was first demonstrated through the simulation of reinforcement learning agents performing this novel task (Kool et al., 2016). These simulations showed that the degree of model-based control was positively associated with average reward rate on the novel two-step paradigm (see Figure 5A). A subsequent experiment provided convergent evidence for this theoretical result. Kool et al. (2016) found that, across participants, the degree of model-based control positively predicted the average reward rate (Figure 5B), and this correlation was significantly stronger than in the Daw two-step task.

Interestingly, Kool et al. (2016) also observed that participants spontaneously increased their rates of model-based control on the novel two-step task compared to the Daw two-step task. This suggests that the existence of the control-demand tradeoff in the novel paradigm may have triggered a shift towards model-based control. Note that this result is consistent with the cost-benefit hypothesis of arbitration between habit and planning. However, alternative explanations are possible. For example, it may be the case that the introduction of negative reward in the novel paradigm triggered a shift towards model-based control, due to loss aversion. Such a shift would be the result of a decision heuristic signaling that certain features of the task should lead to increased model-based control, regardless of whether it actually yield larger overall reward than model-free control.

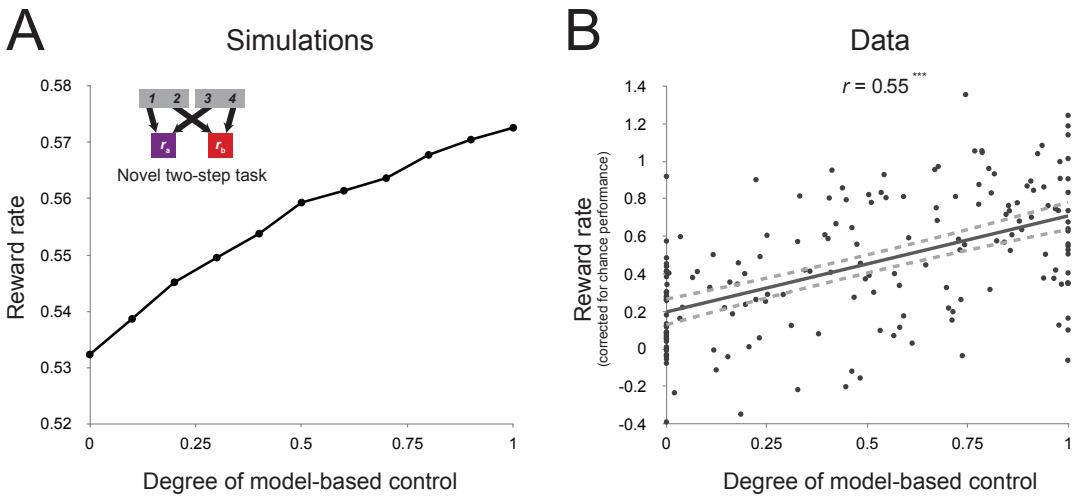

Figure 5. Adapted from Kool et al. (2016). Control-reward tradeoff in the novel two-step task. (A) The relationship between the degree of model-based control and reward rate across 1000 simulations. In contrast with the Daw two-step task, these simulation results show that the novel two-step task successfully achieves a trade-off between model-based control and reward. (B) Relationship between the estimated degree of model-based control and reward rate in the novel two-step task. Consistent with simulation results, there was a strong correlation between these variables (n = 184). Dashed lines indicate the 95% confidence interval. *** $p < 0.001$

Testing the cost-benefit model of arbitration

To distinguish between these two accounts, Kool, Gershman and Cushman (in press) adapted the novel two-step paradigm so that the size of potential reward (the "stakes") changes randomly from trial to trial. In this new task, participants are cued about the size of the stakes at the outset of each trial. The size of the stakes is randomly selected on each trial, with high stakes calculated as a quintupling of baseline rewards. If behavior on this task is determined by a cost-benefit analysis, then people should employ more model-based control in the face of increased incentives, since on those trials the cost-benefit tradeoff would be most beneficial. The results from this experiment were consistent with this hypothesis. Participants showed increased reliance on model-based control on high-stakes trials, indicating an increased willingness to engage in effortful planning (Kool et al., in press).

Even though this result is consistent with the tradeoff hypothesis, it is also consistent with an account which does not rely on the flexible and adaptive integration of costs and benefits. Specifically, participants may have simply acted on a decision heuristic which reflexively increases model-based control in high-stake situations, regardless of whether this provides a reward advantage. To test this possibility, Kool et al. (in press) also implemented the stakes manipulation in the Daw two-step paradigm, since in this task there exists no tradeoff between control and reward. If the stakes effect is driven by an incentive heuristic, high stakes should trigger increased model-based control in both tasks. However, under a cost-benefit account, where the brain estimates task-specific controller values for both systems, model-based control should not increase on high-stakes trials in the stakes version of the Daw two-step task. The results supported the latter hypothesis. Participants who completed the original Daw two-step task were insensitive to reward amplification through the stakes manipulation, in contrast with the increase in model-based control to reward amplification in the novel paradigm (Kool et al., in press).

These results provide the first evidence that the brain attaches a cost to the exertion of model-based control. Furthermore, they provide insight into the way humans arbitrate between control mechanisms. Rather than relying on a heuristic of increasing model-based control when presented with larger incentives or other task features, participants seemed to engage in an adaptive integration of costs and benefits for either strategy in the current environment. Participants flexibly estimated the expected rewards for each system and then weighed this against the increased costs of model-based control.

Alternative models of arbitration

The cost-benefit account of competition between reinforcement learning systems is broadly consistent with two bodies of research. First, the assumption that model-based control carries an intrinsic effort cost finds resonance in a large literature on the aversive nature of cognitive control (Botvinick & Braver, 2015; Gershman et al., 2015; Griffiths et al., 2015; Payne et al., 1993; Rieskamp & Otto, 2006; Shenhav et al., 2013). This work suggests that the exertion of cognitive control can best be understood as the output of cost-benefit analysis. The comparison of behavior between the novel and Daw two-step tasks described above indicates that a similar tradeoff guides the allocation of model-

based control, presumably because this also requires the exertion of cognitive control (Otto, Gershman, et al., 2013; Smittenaar et al., 2013).

Second, there are now several other models of arbitration between competing reinforcement learning system that are, to varying degrees, compatible with the cost-benefit tradeoff account, but which differ in their details (Daw et al., 2005; Keramati et al., 2011; Lee et al., 2014; Pezzulo et al., 2013). Below, we will describe how these models implement the competition between model-free and model-based control and contrast them with our cost-benefit account.

According to Daw et al. (2005), arbitration between conflicting systems for behavioral control is primarily determined on the basis of uncertainty. Specifically, this model estimates each system's value uncertainty for each state-action pair. The model-based system has uncertainty due to bounded computational resources, whereas the model-free system has uncertainty due to limited experience in the environment. These measures of uncertainty are computed through Bayesian implementations of both reinforcement learning systems as the posterior variance of the action values. After estimating these two different forms of uncertainty, the competition is then resolved by choosing the action value of the system with lower uncertainty.

A related metacontrol model uses signals of the systems' reliability as a means of arbitration (Lee et al., 2014). Here, the measure of reliability for a system is proportional to the absolute size of their prediction errors, the degree to which the systems predicted future states or rewards accurately. Similar to the Daw et al. (2005) model, Bayesian estimation of reliability still occurs for the model-based system, while a Pearce-Hall associability-like rule is used to estimate the reliability of the model-free system. In addition, this model also incorporates a 'model bias' term, which favors the model-free system all else being equal, so as to account for differences in cognitive effort. The resulting arbitration process transforms these variables into a weighting parameter, which is then used to compute a weighted combination of action values to guide decision making. Note that, in contrast to the Daw et al. (2005) model, the competition is resolved as a function of the average reliability of the model-based and model-free systems, and not separately for each action.

These closely related models of metacontrol account for many experimental findings, such as the finding that as the model-free system becomes more accurate, agents become increasingly insensitive towards outcome devaluation (since the model-free system needs to incrementally relearn its action-outcome contingencies). Furthermore, the reliability signals in the Lee et al. (2014) model have been shown to have a neural correlate in inferior lateral prefrontal cortex. They cannot, however, explain the observation of increased model-based control on high stakes trials (Kool et al., in press), since the accuracy of either system's prediction does not change as a result of the amplification of reward. Therefore, these models do not predict an increase in proactive model-based control in the face of increased reward potential.

Instead, our cost-benefit hypothesis and the data described above align more strongly with metacontrol models that balance accuracy against control costs. One such model is proposed by Keramati et al. (2011). According to this account, the choice between model-based and model-free control is essentially about maximizing total reward. At each time point, the decision maker estimates the expected gain in reward from running a model-based estimation of action values. This measure, also known as the value of information

(Howard, 1966), was originally developed as a way to negotiate the exploration-exploitation tradeoff in reinforcement learning. Next, the agent also estimates the cost of running those simulations. This cost is explicitly formalized as the amount of potential reward that the model-free system could have accrued while the model-based system is engaged in these prospective simulations. In other words, the cost of model-based control is explicitly an opportunity cost directly proportional to the required processing time. Finally, the costs and gains are compared against each other, and their relative size determines whether the model-based system is invoked. If the costs outweigh the gains, the faster habitual system is employed, otherwise the agent engages in slower model-based planning.

Pezzulo et al. (2013) have developed a related value-based account of arbitration between habit and planning. Similar to the proposal of Keramati et al. (2011), the agent assesses each available action in the current state by first computing the value of information (Howard, 1966) associated with model-based planning. This variable encompasses both the uncertainty about the action's value and also the difference in value between each action and the best available alternative action. The value of information increases when the uncertainty about the current action is high, and also if the difference between possible action values is small (that is, if the decision is more difficult). Next, this measure of the expected gains of model-based control is compared against a fixed threshold that represents the effort cost (Gershman & Daw, 2012) or time cost associated with planning. Again, if the cost exceeds the value of information, the agent relies on cached values, otherwise it will employ model-based simulations over an internal representation of the environment to reduce the uncertainty about the current action values (Solway & Botvinick, 2012).

Both the Keramati et al. (2011) and Pezzulo et al. (2013) models account for a range of behavioral findings. The time-based account of Keramati et al. (2011) model accounts for the increasing insensitivity to outcome devaluation over periods of training. It can also naturally incorporate the finding that response times increase with the number of options, especially early in training, since at those moments the model will engage in time-consuming model-based simulations across the decision tree. Relatedly, Pezzulo et al. (2013) showed that, in a multi-stage reinforcement learning task, their model switches from a large number of model-based simulations in earlier stages towards more reliance on model-free control in later stages. In other words, when the model-free system has generated a sufficiently accurate representation of the world, the agent then prefers to avoid the cost of model-based control. The Pezzulo et al. (2013) model is also able to flexibly shift between systems. For example, it shows a rebalancing towards model-based control in response to a change in reward structure of the environment, i.e., an increase in uncertainty of action outcomes.

However, these models still arbitrate between habit and planning as a function of the amount of uncertainty about value estimates in the model-free action values: both models assume an advantage for model-based control when uncertainty about model-free estimates is high (Keramati et al., 2011; Pezzulo et al., 2013). In doing so, they are not immediately able to explain the effect of increased stakes on model-based control (Kool et al., in press). Those data instead favor a mechanism that directly contrasts the rewards obtained by model-based and model-free control, discounted by their respective cost. Furthermore, the fact that these models require the explicit computation of the expected

gains from model-based simulations (the value of information; Howard, 1966) creates the problem of infinite regress (Boureau et al., 2015). If the purpose of metacontrol is to avoid unnecessary deployment of cognitive control, then this purpose is undermined by engaging in an explicit and demanding computation to determine whether cognitive demands are worthwhile.

Based on the evidence described here, we make two suggestions for new formal models of arbitration between reinforcement learning systems. First, they should incorporate a direct contrast between the costs and benefits of both model-free and model-based learning strategies in their current environment, perhaps in addition to a drive to increase reliability of controller predictions. This property should afford flexible adaptive control in response to the changing potential for reward, such as in the stake-size experiment described above. Second, in order to avoid the issue of infinite regress, the arbitration between habit and planning should be guided by a process that does not involve control-demanding computations of reward advantage, such as the value of information (Howard, 1966). Instead, new models of metacontrol should focus on more heuristic forms of arbitration. Notably, a system that attaches an intrinsic cost to model-based planning might guide meta-control with enhanced efficiency, by circumventing the need for an explicit computation of those costs in terms of effort, missed opportunities, and time. In sum, these properties motivate our proposal that a form of model-free reinforcement learning integrates the reward history and control costs associated with different control mechanisms. The resulting 'controller values' dictate controller arbitration.

Principles of cooperation

While the evidence reviewed in the previous section supports competitive architectures, recent evidence also suggests a variety of cooperative interactions between model-free and model-based reinforcement learning. In this section, we review three different flavors of cooperation.

Model-based simulation as a source of training data for model-free learning

One way to think about the tradeoff between model-free and model-based algorithms is in terms of *sample complexity* and *time complexity*. Sample complexity refers to the number of training examples a learning algorithm needs to achieve some level of accuracy. Time complexity refers to how long an algorithm takes to execute. Intuitively, these correspond to "learning time" and "decision time".

Model-free algorithms have high sample complexity but low time complexity—in other words, learning is slow but deciding is fast. Model-based algorithms have the opposite property: relatively low sample complexity, assuming that the model can be learned efficiently, but high time complexity. Since the amount of data that an agent has access to is typically fixed (by the world, or by the experimenter) and thus beyond algorithmic improvement, it might seem that this tradeoff is inevitable. However, it is possible to create additional examples simply by *simulating* from the model and allowing model-free algorithms to learn from these simulated examples. In this way, the model-based system can manufacture an arbitrarily large number of examples. As a consequence,

the model-free system's sample complexity is no longer tied to its real experience in the world; model-based simulations, provided they are accurate, are a perfectly good substitute.

Sutton (1990) proposed a cooperative architecture called *Dyna* that exploits this idea. A model-free agent, by imbibing model-based simulations, can become arbitrarily proficient without increasing either sample complexity or time complexity. The only requirement is that the agent has sufficient spare time to process these simulations. Humans and many other animals have long periods of sleep or quiet wakefulness during which such simulation could plausibly occur. Notably, neurons in the hippocampus tuned to spatial location ("place cells") replay sequences of firing patterns during rest and sleep (see, Carr, Jadhav, & Frank, 2011 for a review), suggesting they might act as a neural substrate for a Dyna-like simulator (Johnson & Redish, 2005). Furthermore, it is well-known that motor skills can improve following a rest period without additional training (Korman, Raz, Flash, & Karni, 2003; Walker, Brakefield, Morgan, Hobson, & Stickgold, 2002), and reactivating memories during sleep can enhance subsequent task performance (Oudiette & Paller, 2013). Ludvig, Mirian, Kehoe and Sutton (2017) have argued that simulation may underlie a number of animal learning phenomena (e.g., spontaneous recovery, latent inhibition) that are vexing for classical learning theories (which are essentially variants of model-free algorithms).

A series of experiments reported by Gershman, Markman and Otto (2014) attempted to more directly test Dyna as a theory of human reinforcement learning. The experimental design is summarized in Figure 6A. In Phase 1, subjects learn the structure of a simple two-step sequential decision problem. In Phase 2, they learn that taking action A in state 1 is superior to taking action B. They then learn in Phase 3 that state 3 is superior to state 2. This sets up a conflict with what they learned in Phase 2, because taking the preferred action A in state 1 will lead them to state 2 (the inferior state). In Phase 4, Gershman et al. (2014) tested whether they switch their preference for action A following their experience in the second-step states.

Standard model-free learning algorithms like temporal difference learning do not predict any revaluation, because they rely on unbroken trajectories through the state space in order to chain together reward predictions. These trajectories were deliberately broken in the experimental structure so as to handicap model-free learning. Less obviously, standard model-based learning algorithms *also* predict no revaluation, because subjects are explicitly instructed in Phase 4 that they are only being rewarded for their actions in the first state. Thus, the optimal model-based policy should completely ignore information about the second step. Crucially, Dyna predicts a positive revaluation effect, because model-based simulation can effectively stitch together the state sequences which were not explicitly presented to subjects, allowing model-free algorithms to revise the value estimate in state 1 following experience in states 2 and 3.

The experimental results showed clear evidence for a revaluation effect (Figure 6B), supporting the predictions of Dyna. Additional support came from several other findings. First, cognitive load during Phase 3 reduced the revaluation effect. This is consistent with the idea that model-based simulation, like other model-based processes, is computationally intensive and thus susceptible to disruption by competition for resources. Second, the load effect could be mitigated by increasing the number of trials (i.e., opportunities for revaluation) during Phase 3. Third, a brief rest (quiet wakefulness)

prior to Phase 4 increased revaluation, consistent with the hypothesis of offline simulation driving model-free learning (Figure 6C). Finally, applying cognitive load during Phase 4 had no effects on the results, supporting our proposal that performance is driven by model-free control (recall that cognitive load has a selective, deleterious effect on model-based control; Otto, Gershman, et al., 2013).

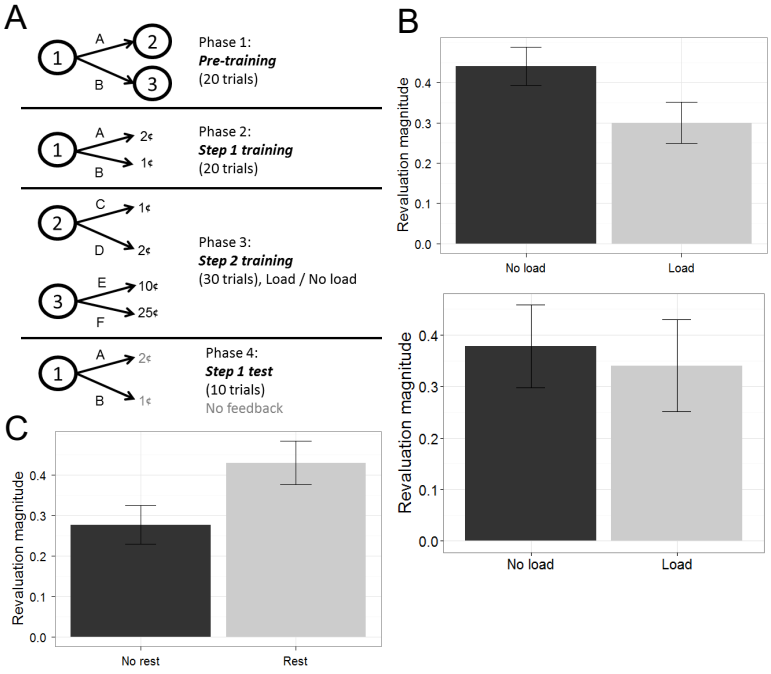

Figure 6. (A) The sequential decision problem consists of three states (indicated by numbered circles) and two mutually exclusive actions in each state (indicated by letters). Deterministic transitions between states conditional upon the chosen action are indicated by arrows. Rewards for each state-action pair are indicated by amounts (in cents). In Phase 4, reward feedback is delayed until the end of the phase. (B) Revaluation in load and no load conditions. Revaluation magnitude is measured as $P_4(\text{action}=B \mid \text{state}=1) - P_3(\text{action}=B \mid \text{state}=1)$, where $P_i(\text{action}=a \mid \text{state}=s)$ is the probability of choosing action a in state s during Phase i. Top: load applied during Phase 3; Bottom: load applied during Phase 4. (C) A brief rest phase prior to Phase 4 ameliorates the effects of load.

Taken together, these results provide some of the first behavioral evidence for cooperative interaction between model-based and model-free reinforcement learning. The same framework may explain the observation that model-based control on the Daw two-step task becomes resistant to disruption by cognitive load over the course of training (Economides, Kurth-Nelson, Lübbert, Guitart-Masip, & Dolan, 2015). If one effect of training is to inject model-based knowledge into the model-free value function, then the model-free system will be able to exhibit model-based behavior autonomously. Dyna may also shed light on the recent observation that dopamine neurons signal prediction errors based on *inferred* (i.e., simulated) values (Doll & Daw, 2016; Sadacca, Jones, & Schoenbaum, 2016).

Partial evaluation

Keramati et al. (2016) have investigated an alternative way to combine model-based and model-free systems, which they refer to as "planning-until-habit", a strategy closely related to "partial evaluation" in the computer science literature (see Daw & Dayan, 2014). The basic idea, illustrated in Figure 7, is to do limited-depth model-based planning, and then insert cached model-free values at the leaves of the decision tree. The sum of these two components will equal the full value at the root node. This model nests pure model-based (infinite depth) and pure model-free (depth 0) algorithms as special cases. The primary computational virtue of partial evaluation is that it can efficiently exploit cached values to augment model-based planning. This will work well when cached values are accurate in some states but not others (where planning is required).

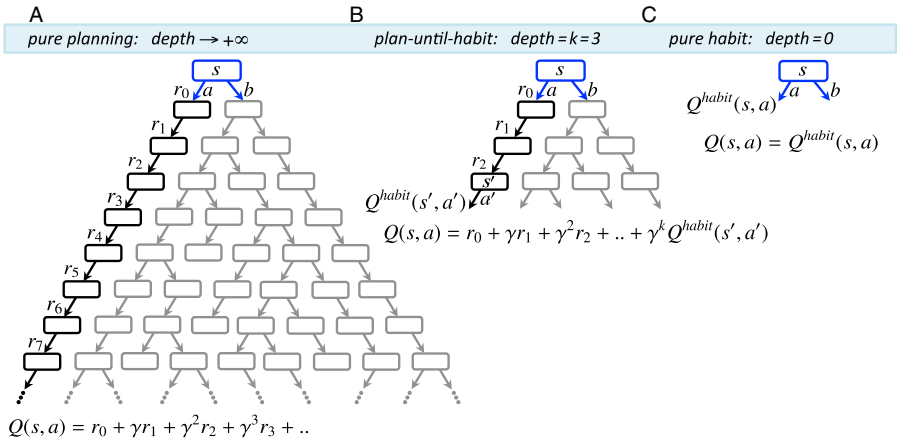

Figure 7. Reprinted from Keramati et al. (2016). (A) Pure planning: rewards are mentally accumulated over an infinite horizon. (B) Plan-until-habit: rewards are partially accumulated, and then combined with a cached value function. (C) Pure habit: actions are evaluated using only cached values, no reward accumulation.

Keramati et al. (2016) provided behavioral evidence for this proposal using a novel three-step extension of the Daw two-step task. Using the same logic of analyzing the interaction between reward outcome and transition probability on subsequent choices, they found differences in the mixture of model-based and model-free behavior at different steps of the task. In particular, subjects appeared model-based with respect to the second step but model-free with respect to the third step, precisely what was predicted by the partial evaluation strategy. Moreover, putting people under time pressure shifted them to a pure model-free strategy at both steps, consistent with the idea that the depth of model-based planning is adaptive and depends on resource availability.

Habitual goal selection

An advantage of model-based control is its capacity to plan towards goals. That is, a model-based agent can specify any particular state of the world that she wishes to attain

(e.g., being at the dentist's office at 2pm with a bottle of ibuprofen), and then evaluate candidate policies against their likelihood of attaining that goal state. In many laboratory tasks, the number of possible goal states may be very small, or they may be explicitly stated by the experimenter. For instance, in the classic "two-step" task presented in Figure 1, there are only six states towards which the agent might plan (two intermediate states and four terminal states). In the real world, however, the number of possible goal states that we might select at any given moment is very large. Usually there are no experimenters restricting this set for you. How do we decide which goals to pursue?

One possibility is exhaustive search, but this is computationally prohibitive. Consider, for instance, evaluating candidate goals alphabetically: You could set the goal of *absconding with an aardvark*, or *absconding with an abacus*, and so on, until eventually considering selecting the goal of *x-raying a xerox*. For the same reason--i.e., the large set of possible goals in most real-world settings--it is not practical to employ model-based evaluation of the rewards of candidate goals in order to decide which goal to select. Is there a more efficient way to decide which particular goal to pursue from moment to moment?

An obvious alternative is to select goals by model-free methods--in other words, to store a state-specific cached value of the likely value of pursuing different goals. Put simply, an agent might ask herself, "when I've been in this situation in the past, what have been rewarding goals for me to select?" Of course, once a goal is selected, it falls to model-based processes to plan towards that goal. This entails a cooperative relationship between the two control mechanisms: Cached, model-free values may be used to decide *which goal to pursue*, while model-based planning is used in order to determine *how to attain it*.

The utility of this approach is best appreciated through a specific example (Cushman & Morris, 2015). Consider an experienced journalist who sets out to report on different news events each day. At a high level of abstraction her job is structured around a regular series of goals to pursue: "Find out what has happened this morning"; "Consult with my editor"; "Obtain interviews"; "Write a draft", and so forth. Thus, selecting goals may be efficiently accomplished by consider their cached value: "Obtaining interviews" was a valuable goal yesterday, and it will remain so today. Yet, pursuing any one of these goals would require flexible model-based planning--for instance, the actions necessary to interview the president one day will be different than the actions necessary to interview a political dissident the next day. In sum, then, a favorable architecture for many tasks would select goals according to model-free value, but then attains goals by model-based planning.

Cushman and Morris (2015) found empirical support for this architecture using several modified versions of the classic Daw two-step task. An example is illustrated in Figure 8. The essence of the design is to prompt people to choose an action that reveals their goal, but then occasionally transition them to a non-goal state. If this reinforcement history affects their subsequent choice despite its low probability, then it can be attributed to a model-free value update process. Subsequently participants are tested on different actions that are associated with common goal states. Influence of reinforcement history even upon these different actions implies a model-free value assignment not to the action itself, but rather to the goal state with which it is associated.

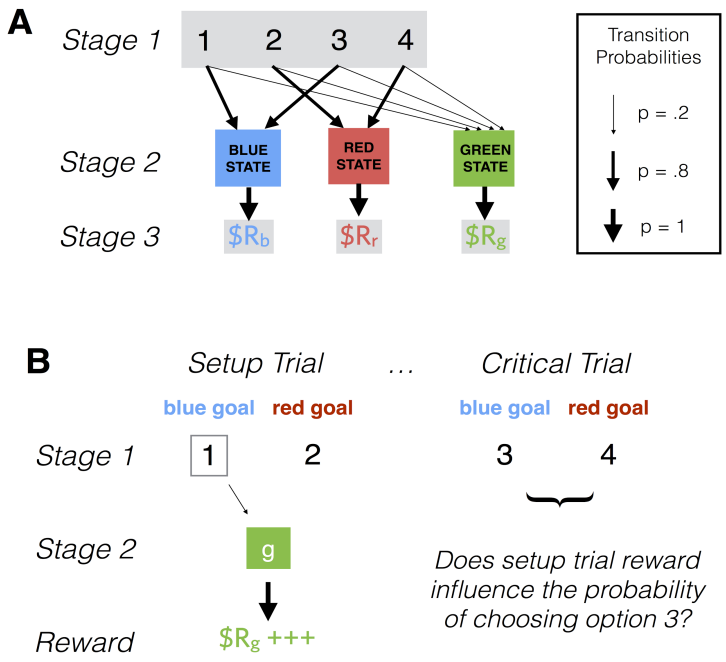

Figure 8. Reprinted from Cushman & Morris (2015). A modified version of the two-step task designed to test a model of habitual goal selection. (A) At stage 1 participants are presented with two available actions drawn from a set of four (1, 2, 3 and 4). These transitions with high probability to either a blue or red intermediate state, and with equal low probability to a green state. (B) On critical trials, the low-probability green state transition occurs. The key question is whether the reward obtained following the green state influences subsequent choice of different actions that share the same goal (e.g., whether a reward following the sequence 1, Green influences the probability of subsequently choosing action 3, which shares the blue stage goal with action 1). Across several experiments, participants exhibited precisely this effect.

Beyond the particular case of goal selection, this research points towards a more general form of cooperative interaction between model-free and model-based systems. For typical real-world problems, full model-based evaluation of all possible action sequences will always pose prohibitive computational demands. One solution to this problem is to use cached, model-free values to weight the probability with which all possible actions are introduced into the subset of actions that receive further model-based evaluation. (This subset might be described as the "choice set"). All else being equal, the higher the cached value of an action, the more likely that the benefits of a more precise model-based estimate of its current value will outweigh the computational demands involved. Investigating this general process of "choice set construction" is an important direction for future research.

Conclusion

Over the last century, the idea that human behavior is controlled by two systems, one habitual and one goal-directed, has become a cornerstone of psychological and behavioral theories of cognition and decision making (Dickinson, 1985; Dolan & Dayan, 2013; Fudenberg & Levine, 2006; Kahneman, 1973; Sloman, 1996). Recent reinforcement learning theory has brought mathematical precision to this area of research by formalizing

this distinction in terms of model-based and model-free control (Daw et al., 2011; Daw et al., 2005; Gläscher et al., 2010). We have reviewed the surge of empirical and theoretical research emanating from this formalism.

First, we reviewed work that addresses how the habitual and goal-directed systems are engaged in a competition for control of behavior. We proposed that this competition is arbitrated as a tradeoff between the costs and benefits of employing each system. At the core of this proposal is the idea that the exertion of model-based control carries an intrinsic effort cost associated with the exertion of cognitive control. This account is supported by the findings that model-based planning is dependent on cognitive resources (Otto, Gershman, et al., 2013; Otto, Raio, et al., 2013; Otto et al., 2015; Schad et al., 2014), and that humans attach intrinsic disutility to the exertion of cognitive control (Kool et al., 2010; Westbrook et al., 2013). Current research indicates that model-based control is spontaneously increased in response to reward amplification, but only when the model-based system is associated with increased accuracy (Kool et al., 2016; Kool et al., in press). Together, these findings suggest that the brain estimates values for each system, integrating their costs and benefits into a single metacontrol value that it uses to guide controller arbitration.

Second, we reviewed a new line of research that focuses on the ways in which habit and planning act in a cooperative fashion to achieve both efficiency and accuracy. Evidence suggests a plethora of cooperative strategies: the model-free system can learn from data simulated from the model-based system (Gershman et al., 2014), can truncate model-based planning (Keramati et al., 2016) or can facilitate the selection of rewarding goals (Cushman & Morris, 2015). At present, it is unclear whether these different strategies occur simultaneously, or are adaptively invoked much like in the controller arbitration problem.

In the work described here, the idea of an intrinsic effort cost for model-based control has only come to the fore in the research on the competitive interaction between habit and planning. However, given the ubiquitous nature of the cost for cognitive control (Botvinick & Braver, 2015; Westbrook & Braver, 2015), such a cost is likely to also play a role in the collaborative interactions between these two systems. From this perspective, several intriguing questions arise.

Some of these questions concern the basic algorithmic approach that the brain takes to decision-making. For instance, is habitual goal selection (Cushman & Morris, 2015) more prevalent for people who attach a higher intrinsic cost to model-based planning? Does the intrinsic cost of cognitive control establish the threshold at which estimation of action values switches from planning to habit in the situations described by (Keramati et al., 2016)? In light of our cost-benefit theory of controller arbitration, one may view the cooperative interaction between habit and planning as a case of bounded rationality (Gigerenzer & Goldstein, 1996). From this perspective, costly cognitive resources would be deployed to maximize accuracy among a restricted region of the action space while preserving a net gain in value, and habit would provide complementary assistance for those actions not analyzed through model-based control. Note that this framework predicts that increased potential incentives (as used in Kool et al., in press), will lead to deeper planning in the Keramati et al. (2016) task, and a reduced reliance on habitual goal selection in the Cushman and Morris (2015) task.

Other questions involve neural implementation. Ever since the recent resurgence of reinforcement learning theory in modern psychological research, the neuromodulator dopamine has come to the fore as playing a key role. Most famously, Schultz, Dayan and Montague (1997) showed that reward prediction errors, the signals that drive learning of action-outcome contingencies, are encoded by the phasic firing of dopamine neurons that project to the ventral striatum in the basal ganglia. More important for the current purpose, it has been suggested that tonic levels of dopamine encode an average reward signal that determines response vigor in operant conditioning tasks (Hamid et al., 2016; Niv, Daw, & Dayan, 2006), so higher dopamine levels yielding increased responding on free-operant conditioning tasks. Based on these and related results, Salamone and colleagues (Salamone & Correa, 2012; Salamone, Correa, Farrar, Nunes, & Pardo, 2009) have proposed that baseline levels of dopamine in the basal ganglia may actually serve to discount the perceived costs of physical effort. For example, rats in an effort-based decision-making task, show reduced willingness to climb over barriers to obtain rewards after depletion of dopamine in the nucleus accumbens (Cousins, Atherton, Turner, & Salamone, 1996). Westbrook and Braver (2016) have proposed a very similar view for the case of mental effort. According to this account, increases in baseline dopamine levels in response to high-reward situations facilitate subsequent cognitive processing by enhancing stability of working memory representations in prefrontal cortex. Intriguingly, recent experiments indicate that baseline dopamine levels in the ventral striatum correlated positively with a bias towards more model-based control (Deserno et al., 2015) and that experimentally induced increases in dopamine increase the degree of model-based control in the Daw two-step task (Sharp, Foerde, Daw, & Shohamy, 2015; Wunderlich, Smittenaar, & Dolan, 2012; see Chapter 11 by Sharpe & Schoenbaum in this volume). Together, these insights hint at the intriguing possibility that this effect of dopamine on model-based control may be viewed as the result of an alteration of the variables that enter the cost-benefit tradeoff at the algorithmic level.

While the work we have reviewed in this chapter suggests a rich space of competition and cooperation between reinforcement learning systems, we have in fact only skimmed the surface. New research suggests separate but interacting systems for Pavlovian (Dayan & Berridge, 2014) and episodic (Gershman & Daw, 2017) reinforcement learning. One may reasonably worry that theorists are gleefully manufacturing theories to accommodate each new piece of data, without addressing how the systems act in concert as part of a larger cognitive architecture. What is needed is a theory of metacontrol that encompasses all of these systems. The development of such a theory will be a central project for the next generation of reinforcement learning research.

References

Akam, T., Costa, R., & Dayan, P. (2015). Simple plans or sophisticated habits? State, transition and learning interactions in the two-step task. *PLoS Computational Biology, 11*, e1004648.

Army Individual Test Battery. (1944). *Manual of Directions and Scoring*. Washington, DC: War Department, Adjutant General's Office.

Balleine, B. W., & O'Doherty, J. (2009). Human and rodent homologies in action control: Corticostrialtal determinants of goal-directed and habitual action. *Neuropsychopharmacology, 35*, 48-69.

Botvinick, M. M., & Braver, T. (2015). Motivation and cognitive control: From behavior to neural mechanism. *Annual Review of Psychology, 66*, 83-113.

Boureau, Y.-L., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding how to decide: Self-control and meta-decision making. *Trends in Cognitive Sciences, 19*, 700-710.

Carr, M. F., Jadhav, S. P., & Frank, L. M. (2011). Hippocampal replay in the awake state: a potential physiological substrate of memory consolidation and retrieval. *Nature Neuroscience, 14*, 147-153.

Cousins, M. S., Atherton, A., Turner, L., & Salamone, J. D. (1996). Nucleus accumbens dopamine depletions alter relative response allocation in a T-maze cost/benefit task. *Behavioural brain research, 74*, 189-197.

Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Science, 112*, 13817–13822.

Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*, 20130478.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*, 1204-1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*, 1704-1711.

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience, 14*, 473-492.

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological Science, 27*, 848-858.

Deserno, L., Huys, Q. J. M., Boehme, R., Buchert, R., Heinze, H.-J., Grace, A. A., . . . Schlagenhauf, F. (2015). Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences, 112*, 1595-1600.

Dezfouli, A., & Balleine, B. W. (2013). Actions, action sequences and habits: Evidence that goal-directed and habitual action control are hierarchically organized. *PLoS Computational Biology, 9*, e1003364-1003314.

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences, 308*, 67-78.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron, 80*, 312-325.

Doll, B. B., & Daw, N. D. (2016). The expanding role of dopamine. *eLife, 5*, e15963.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience, 18*, 767-772.

Economides, M., Kurth-Nelson, Z., Lübbert, A., Guitart-Masip, M., & Dolan, R. J. (2015). Model-based reasoning in humans becomes automatic with training. *PLoS Computational Biology, 11*, e1004463-1004419.

Fudenberg, D., & Levine, D. K. (2006). A dual self model of impulse control. *American Economic Review, 96*, 1449-1476.

Gershman, S. J. (2017). Reinforcement learning and causal models. In M. Waldmann (Ed.), *Oxford Handbook of Causal Reasoning*: Oxford University Press.

Gershman, S. J., & Daw, N. (2012). Perception, action and utility: the tangled skein. In M. Rabinovich, K. Friston, & P. Varona (Eds.), *Principles of Brain Dynamics: Global State Interactions*. Cambridge, MA: MIT Press.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annual Review of Psychology, 68*, 101-128.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science, 349*, 273-278.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General, 143*, 182-194.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review, 103*, 650-669.

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife, 5*, e11305.

Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience, 15*, 523-536.

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

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science, 7*, 217-229.

Hamid, A. A., Pettibone, J. R., Mabrouk, O. S., Hetrick, V. L., Schmidt, R., Vander Weele, C. M., . . . Berke, J. D. (2016). Mesolimbic dopamine signals the value of work. *Nature Publishing Group, 19*, 117-126.

Howard, R. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics, 2*.

Johnson, A., & Redish, A. D. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Networks, 18*, 1163-1171.

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology, 7*, e1002055.

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. *Proceedings of the National Academy of Sciences of the United States of America, 113*, 12868-12873.

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLoS Computational Biology, 12*, e1005090.

Kool, W., Gershman, S. J., & Cushman, F. A. (in press). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*.

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General, 139*, 665-682.

Korman, M., Raz, N., Flash, T., & Karni, A. (2003). Multiple shifts in the representation of a motor sequence during the acquisition of skilled performance. *Proceedings of the National Academy of Sciences, 100*, 12492-12497.

Kurzban, R., Duckworth, A. L., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences, 36*, 661-726.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron, 81*, 687-699.

Ludvig, E. A., Mirian, M. S., Kehoe, E. J., & Sutton, R. S. (2017). Associative learning from replayed experience. *bioRxiv*.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Reviews in Neuroscience, 24*, 167-202.

Niv, Y., Daw, N., & Dayan, P. (2006). How fast to work: Response vigor, motivation and tonic dopamine. *Advances in neural information processing systems, 18*, 1019.

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science, 24*, 751-761.

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E., & Daw, N. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences USA, 110*, 20941-20946.

Otto, A. R., Skatova, A., Madlon-Kay, S., & Daw, N. D. (2015). Cognitive control predicts use of model-based reinforcement learning. *Journal of Cognitive Neuroscience, 27*, 319-333.

Oudiette, D., & Paller, K. A. (2013). Upgrading the sleeping brain with targeted memory reactivation. *Trends in Cognitive Sciences, 17*, 142-149.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 534-552.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge, England: Cambridge University Press.

Pezzulo, G., Rigoli, F., & Chersi, F. (2013). The Mixed Instrumental Controller: Using Value of Information to combine habitual choice and mental simulation. *Frontiers in Psychology, 4*, 92.

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135*, 207-236.

Sadacca, B. F., Jones, J. L., & Schoenbaum, G. (2016). Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *eLife, 5*, e13665.

Salamone, J. D., & Correa, M. (2012). The mysterious motivational functions of mesolimbic dopamine. *Neuron, 76*, 470-485.

Salamone, J. D., Correa, M., Farrar, A. M., Nunes, E. J., & Pardo, M. (2009). Dopamine, behavioral economics, and effort. *Frontiers in Behavioral Neuroscience, 3*, 2-12.

Schad, D. J., Jünger, E., Sebold, M., Garbusow, M., Bernhardt, N., Javadi, A.-H., . . . Huys, Q. J. M. (2014). Processing speed enhances model-based over model-free reinforcement learning in the presence of high working memory functioning. *Frontiers in Psychology, 5*, 1450.

Schouppe, N., Ridderinkhof, K. R., Verguts, T., & Notebaert, W. (2014). Context-specific control and context selection in conflict tasks. *Acta Psychologica, 146*, 63-66.

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

Sebold, M., Deserno, L., Nebe, S., Nebe, S., Schad, D. J., Garbusow, M., . . . Huys, Q. J. M. (2014). Model-based and model-free decisions in alcohol dependence. *Neuropsychobiology, 70*, 122-131.

Sharp, M. E., Foerde, K., Daw, N. D., & Shohamy, D. (2015). Dopamine selectively remediates 'model-based' reward learning: a computational approach. *Brain, 139*, 355-364.

Shenhav, A., Botvinick, Matthew M., & Cohen, Jonathan D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron, 79*, 217-240.

Skatova, A., Chan, P. A., & Daw, N. D. (2015). Extraversion differentiates between model-based and model-free strategies in a reinforcement learning task. *Frontiers in Human Neuroscience, 7*, 525.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3-22.

Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D., & Dolan, R. J. (2013). Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron, 80*, 914-919.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychological Review, 119*, 120-154.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662.

Sutton, R. S. (1990). First results with Dyna, an interesting architecture for learning, planning, and reacting. In T. Miller, R. S. Sutton, & P. Werbos (Eds.), *Neural Networks for Control* (pp. 179–189). Cambridge, MA: MIT Press.

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

Thorndike, E. L. (1911). *Animal Intelligence: Experimental Studies*. New York: The Macmillan Company.

Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A., & Stickgold, R. (2002). Practice with sleep makes perfect: sleep-dependent motor skill learning. *Neuron, 35*, 205-211.

Westbrook, A., & Braver, T. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive Affective, & Behavioral Neuroscience, 15*, 395-415.

Westbrook, A., & Braver, T. S. (2016). Dopamine does double duty in motivating cognitive effort. *Neuron, 89*, 695-710.

Westbrook, A., Kester, D., & Braver, T. S. (2013). What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PLoS ONE, 22*, e68210.

Wunderlich, K., Smittenaar, P., & Dolan, R. (2012). Dopamine enhances model-based over model-free choice behavior. *Neuron, 75*, 418–424.