may have trapped these plants in an adaptive peak. Despite this, there are examples of plant species that have evolved other modes of pollination derived from buzz-pollinated ancestors. For example, a few species of Melastomataceae have evolved flowers with modified anthers that act as bellows to release pollen. When a pollinator collects floral perfumes (non-buzzing male orchid bees in *Solanum*) or remove the sugar-rich structures at the base of modified stamens (birds in Melastomataceae), the flexible anther is squeezed, producing a puff of air, which fires pollen onto the pollinator's body. Having lost their need for buzz pollinators, these flowers demonstrate the diversity of ways in which plants evolve to disperse their pollen.

Where can I find out more?

Buchmann, S.L. (1983). Buzz pollination in angiosperms. In Handbook of Experimental Pollination Biology, C.E. Jones and R.J. Little, eds. (New York, NY: Van Nostrand Reinhold Company), pp. 73–113.

Danforth, B.N., Minckley, R.L., Neff, J.L., and Fawcett, F. (2019). The Solitary Bees: Biology, Evolution, Conservation (Princeton, NJ: Princeton University Press).

Dellinger, A.S., Chartier, M., Fernández-Fernández, D., Penneys, D.S., Alvear, M., Almeda, F., Michelangeli, F.A., Staedler, Y., Armbruster, W.S., and Schönenberger, J. (2019). Beyond buzz-pollination–departures from an adaptive plateau lead to new pollination syndromes. New Phytol. *221*, 1136–1149.

De Luca, P.A., Bussière, L.F., Souto-Vilaros, D., Goulson, D., Mason, A.C., and Vallejo-Marín, M. (2013). Variability in bumblebee pollination buzzes affects the quantity of pollen released from flowers. Oecologia *172*, 805–816.

Harder, L.D., and Barclay, R.M.R. (1994). The functional significance of poricidal anthers and buzz pollination: controlled pollen removal from *Dodecatheon*. Funct. Ecol. *8*, 509–517.

King, M.J., Buchmann, S.L., and Spangler, H. (1996). Activity of asynchronous flight muscle from two bee families during sonication (buzzing). J. Exp. Biol. *199*, 2317–2321.

King, M.J., and Buchmann, S.L. (2003). Floral sonication by bees: mesosomal vibration by *Bombus* and *Xylocopa*, but not *Apis* (Hymenoptera: Apidae), ejects pollen from poricidal anthers. J. Kans. Entomol. Soc. *76*, 295–305.

Switzer, C.M., Russell, A.L., Papaj, D.R., Combes, S.A., and Hopkins, R. (2019). Sonicating bees demonstrate flexible pollen extraction without instrumental learning. Curr. Zool. *65*, 425–436.

Vallejo-Marín, M. (2019). Buzz pollination: studying bee vibrations on flowers. New Phytol. *224*, 1068–1074.

Whitehorn, P.R., Wallace, C., and Vallejo-Marin, M. (2017). Neonicotinoid pesticide limits improvement in buzz pollination by bumblebees. Sci. Rep. *7*, 15562.

Biological and Environmental Sciences,
University of Stirling, Stirling FK9 4LA,
Scotland, UK.
*E-mail: david.pritchard@stir.ac.uk

Primer

Model-based decision making and model-free learning

Nicole Drummond and Yael Niv

Free will is anything but free. With it comes the onus of choice: not only what to do, but which inner voice to listen to — our 'automatic' response system, which some consider 'impulsive' or 'irrational', or our supposedly more rational deliberative one. Rather than a devil and angel sitting on our shoulders, research suggests that we have two decision-making systems residing in the brain, in our basal ganglia. Neither system is the devil and neither is irrational. They both have our best interests at heart and aim to suggest the best course of action calculated through rational algorithms. However, the algorithms they use are qualitatively different and do not always agree on which action is optimal. The rivalry between habitual, fast action and deliberative, purposeful action is an ongoing one.

In this primer, we first expand on the theoretical and computational basis of each of the two systems and their neural substrates. Given the evidence that the two systems work in tandem, we then discuss how the brain arbitrates between them by asking what are the conditions under which each system is preferentially used. Finally, we discuss a task developed to test the theoretical assumptions about the two modes of decision making in humans.

Goal-directed *versus* habitual behavior

The idea of two systems of control, one dependent on deliberation and one based on automatic reactions, has roots in the psychological literature on instrumental conditioning and the differentiation between goal-directed and habitual behavior. In instrumental conditioning, animals learn which action (or series of actions) will bring about a desired outcome, for example rats learning to press a lever for food reward. Outcomes that increase the performance of an action are 'reinforcers' — whether appetitive, in the rat example food increasing lever-pressing, or aversive, as when a shock increases escape behavior.

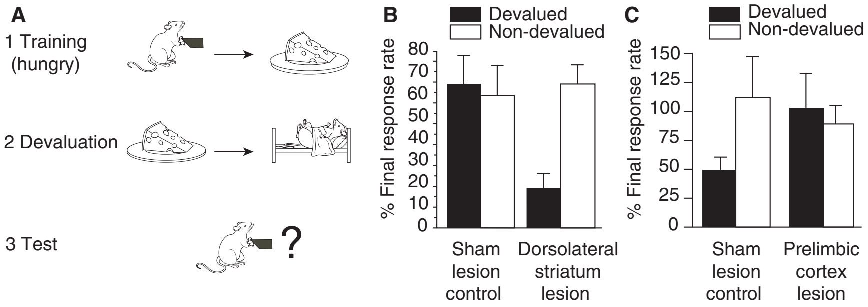
What do animals learn in instrumental conditioning? Early theorists suggested two main alternatives: according to Edward Thorndike, the core of instrumental learning is the formation of associations between stimuli and responses ('S–R' learning), with the reinforcer 'stamping in' these associations, but not being part of the learned construct. In contrast with this behaviorist view, Edward Tolman, foreman of the cognitive revolution, argued that animals learn more elaborate cognitive maps. In these, the learned association is between actions and their outcomes, broadly construed ('A→O' learning), and the association is predicated on their causal contingency (which we denote by →).

As highlighted by Robert Rescorla, Ruth Colwill, Anthony Dickinson and Bernard Balleine, these competing models make contradictory predictions: Tolman's A→O model suggests that behavior should be sensitive to two things: changes in the value of the outcome, for example, if the outcome is no longer desired due to satiety, the rat will decrease its rate of pressing; and changes in the contingency between the action and the outcome, for example, if the action no longer leads to the outcome, the rat will cease to perform the action.

Thorndike's theory, on the other hand, predicts that learned actions will be impervious to such environmental (or internal) changes, as once they are stamped in, they will not be 'stamped out' except by aversive outcomes. Indeed, not all decisions to press a lever are created equal: for an animal to make a deliberative, purposeful choice, it must choose the action because it knows it will result in a food pellet, and the food pellet is desirable to it (as in Tolman's model). However, as per Thorndike, this is not the only reason an animal may choose to press a lever.

What do animals do when the environment changes in such ways? It depends. In particular, early findings showed that rats that were trained extensively to press the lever indeed persisted in pressing for an undesirable outcome, as predicted by stimulus–response learning. However, rats that were only trained for a moderate amount decreased their lever-pressing substantially if the outcome of a lever-press was devalued, consistent with action→outcome learning (Figure 1). Indeed, the distinction between 'goal-directed' and 'habitual' instrumental actions hinges on whether said actions maintain adaptability in response to changes in outcome and/or contingency is considered, or are impervious to such changes, respectively. Mirroring these two behavioral patterns, there is evidence for two separate systems for action control in the brain, as we will detail below.

Humans, too, show both habitual and goal-directed modes of action: after moving house, we can flexibly plan a new route home from work; however, if we are talking to a friend, or listening to the radio, we might habitually make the wrong turn to our old address instead. As in rats, habits are bred through repetition — but they are not context independent or 'irrational'. As stimulus–response associations, habits can be exquisitely context dependent, that is, dependent on local stimuli. You can walk home habitually, but you will turn left at the correct intersection. These actions have habitized precisely because they have consistently led to favorable outcomes in the past. Further, as the example here suggests, goal-directed action requires more cognitively effortful deliberation, and action selection may give way to habits if those cognitive resources are otherwise occupied, as we discuss below.



Current Biology

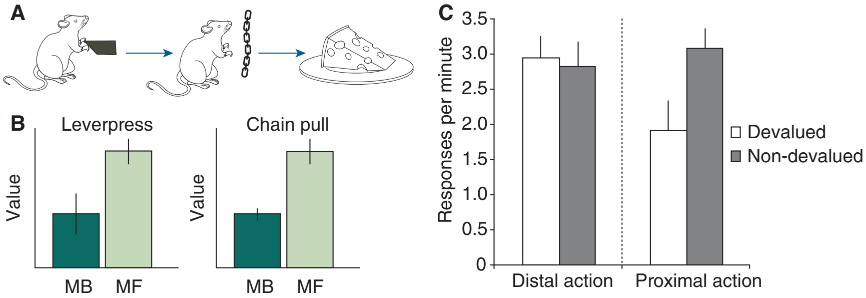
Figure 1. Outcome-devaluation test of goal-directed *versus* habitual control.
(A) In phase 1, rats are trained to press a lever (the prototypical instrumental action) in order to obtain food. Training may be for a short duration (for example, four daily sessions) or extensive (15 sessions). In phase 2, the outcome is devalued, for instance by injecting lithium chloride after consumption, which causes the animal to feel ill. The effects of the manipulation are then tested in a third phase by assessing how eagerly rats will press the lever. In phase 3, the test is performed in extinction (without new outcomes) so as to assess only the products of previous learning: will the value of lever pressing have changed, given the experience in the second phase? Habitual control predicts that it will not (as neither the stimulus nor the response from training were experienced in the devaluation phase, and therefore the stimulus–response association should be intact). In contrast, if the animal is using goal-directed control, it will realize that the action leads to a no-longer-valuable outcome, and therefore the action is moot. Figure adapted from Dickinson and Balleine (1994). (B) After extensive training (in the first phase), rats continue pressing the lever even after devaluation (sham lesion control). However, lesions of the dorsolateral striatum reinstate goal-directed control (lesion group). Blocking dopamine transmission to the dorsolateral striatum during training also shows a similar pattern (not shown here). Plotted is the rate of lever pressing at test, as compared to the last session of training. Figure adapted from Yin *et al.* (2005). (C) In contrast, after only moderate training, rats' lever pressing still diminishes after devaluation (sham lesion control). This devaluation-sensitivity, however, requires an intact prelimbic cortex (lesion group), as well as other brain areas comprising the same cortico-basal-ganglia loop (not shown). Figure adapted from Killcross and Coutureau (2003).

Model-based *versus* model-free reinforcement learning

Machine learning theory offers another, computational, framework for understanding these two decision-making systems. Specifically, Nathaniel Daw and colleagues used the framework of reinforcement learning to provide a formal description of the habitual *versus* deliberative choice mechanisms in terms of model-based and model-free action selection. Reinforcement learning refers to the interaction between a learning agent and its environment in terms of decision making (that is, action selection) within a Markov Decision Process. The Markov Decision Process is a description of the environment as a set of states that transition between each other probabilistically (possibly dependent on an agent's actions), and a set of rewards that may be available in each state. Most reinforcement learning algorithms predicate action choices on 'values': the expected sum of future rewards if a specific action were to be taken at a specific state.

In model-based reinforcement learning, action values are computed by mentally simulating the consequences of said actions within a representational map of the world that includes the states, the transition probabilities between them, and the rewards in each state — a 'world model', hence model-based reinforcement learning. The decision-making agent can learn a model of the environment from experience, by observing and keeping track of sequences of states, and the rewards available in each state, for instance, implemented through learning from state prediction errors. To decide what action to take, the agent can then search through different trajectories of the mental map — 'planning' or mental simulation in a 'forward model' — and choose a path that yields the best estimated long-run outcome.

A benefit of making decisions based on a stored mental map is that the map can easily be updated when changes in the environment occur, meaning that actions that are computed using the map will also flexibly change as needed. So if, for example, the agent learns that a particular route home is unexpectedly blocked, it can immediately re-plan accordingly. However, searching through all possible action trajectories before making a choice is computationally costly, and in some situations, prohibitively so. Such planning and re-planning also takes time, which may not be available for certain decisions that are time-sensitive. This model-based method of action selection maps directly to Tolman's action→outcome model of learning: the world model includes the contingency relationships between actions and outcomes in the state transitions, as well as the subjective worth of the different outcomes. Planning within the model therefore realizes Anthony Dickinson's goal-directed combination of contingency and outcome information in order to make a decision.



Current Biology

Figure 2. Uncertainty-based arbitration between model-free and model-based control.
(A) The sequence of actions required from the rats in Balleine *et al*. (1995). (B) Each system estimates the value of pressing the lever, as well as its own confidence in that value (error bar). Here, after devaluation, the model-free system has more confidence in its value of pressing the lever, while the model-based system is more accurate in evaluating the utility of chain pulling. (C) As a result, after devaluation rats press the lever (action distal from reward) but refrain from subsequently pulling the chain (action proximal to reward). Figure adapted from Balleine *et al.* (1995).

In contrast, in model-free reinforcement learning, decisions are made without a world model. Instead, values for different actions are learned directly, through trial-and-error interaction with the environment every time an action is performed. Decisions then rely on those values that embody a cached summary of past experience. In particular, on each trial, the agent observes the current state, performs an action based on the estimated values of different actions at that state, and observes the results: the outcome it receives, and the state it transitions to. The agent then computes a reward prediction error: the difference between the outcomes obtained and the amount of expected reward (the original estimated value of the action taken). Importantly, because values reflect expected future rewards (not just the immediate reward), to estimate the outcomes obtained, the agent sums up the immediate reward and the value of the new state (the rewards expected from that state onward). The prediction error is then used to determine the magnitude and direction of an update of the value of the action taken: if the prediction error is positive (the action yielded better outcomes than expected), the action value is increased, and *vice versa* if the prediction error is negative.

Through this process of updating values trial-by-trial, prediction errors are minimized and the scalar value of each action eventually approximates the true future expected rewards. Action selection based on model-free reinforcement learning is therefore optimal once learning converges. However, because the values reflect the accumulation of previous rewards and transitions, this method of learning is less flexible in adapting to abrupt change. Known changes in the environment — for instance, hearing about the road block — cannot be incorporated into learned values without directly experiencing prediction errors. This is not just an algorithmic quirk: because cached values embody long-term expected rewards, that is, expected sums of rewards in future states, it is not possible to cleave out the portion of the value that should be changed given a change in one of the state contingencies or predicted outcomes. The information that would allow such dissection — the breakdown of the different components of the total value — is simply not stored in this type of algorithm. Model-free action selection therefore formalizes Thorndike's stimulus-response model: responding is driven by the learned value of a response at a specific state, without explicating the expected outcome.

According to the computational theory, the main benefit of model-free decision making is that action selection is computationally easy: one can simply choose the action with the highest cached value. However, a large amount of trial-and-error experience is required for values to become good estimates of future rewards. Furthermore, as mentioned, the cached values are inflexible and trial-and-error learning is needed to update them. Model-based decision making is therefore advantageous in its adaptability in response to changes in outcome values or state transitions. But it is either computationally costly because of the search over trajectories of possible action sequences, or inaccurate if not all sequences are mentally simulated. As such, neither system should dominate in *all* situations, and if the brain uses the systems optimally, each is expected to have a role to play in decision making.

As an aside, there are additional decision-making algorithms that the brain can — and likely does — employ. For instance, one can select actions by sampling distinct episodic memories (in machine learning this would be akin to sampling a past trajectory and repeating it, with the sampling potentially prioritizing trajectories that led to more rewards). This algorithm would be most appropriate when very few experiences with a task exist, and both model-based and model-free reinforcement learning cannot reliably estimate action values.

Uncertainty-based arbitration between the two systems

If the brain has two mechanisms to estimate the value of different actions, they may not always agree. This then creates a new decision-making problem: which system to use? Experiments testing for the behavioral hallmarks of goal-directed *versus* habitual behavior reveal conditions under which each strategy is employed: early in learning, devaluation results confirm that behavior is goal directed, that is, consistent with a model-based reinforcement learning strategy for computing action values. But once behavior is repeated more extensively ('overtrained'), it turns habitual (see control groups in Figure 1B,C). This is consistent with the idea that model-free learning is slower and less efficient than model-based learning, and so the model-free system should be used only after considerable experience with the situation.

There are also conditions under which behavior seems to remain model-based (that is, sensitive to outcome devaluation) even through extensive training. In particular, when choosing between several different actions that lead to different outcomes — for example, two levers, one leading to delivery of cherry-flavored food pellets and the other to banana-flavored pellets — behavior seems to remain goal-directed for longer. This may be because the choice between different (fairly equivalent) outcomes halves the experience with each action (so there are fewer updates to the model-free value of that action) or because the choice between two actions that are equally favorable in terms of model-free values utilizes the more accurate forward-modeling model-based system in an attempt to disambiguate the options.

One thing that these results show — and that other findings confirm — is that animals can alternate between using each of the decision-making systems within a given local scenario. We are not habitual *or* goal directed, we can be either at any point in time, and we can switch between decision-making modes even within a chain of actions. This was most starkly demonstrated by Balleine and colleagues in an experiment where rats were required to perform a sequence of actions consisting of a lever-press followed by a chain pull in order to obtain a reward outcome (Figure 2A). Animals were trained on the sequence when hungry, but then tested (in extinction) when sated, as a means to devalue the food reward. The results showed that the action distal from the reward (the lever press) was not sensitive to the motivational shift, suggesting habitual control, whereas the action more proximal to the outcome was sensitive to the lower value of the reward, supporting goal-directed control (Figure 2C). That is, the rats continued to press the lever, but substantially reduced their tendency to continue to pull on the chain.

One explanation for this is that proximity to the outcome allows more accurate forward-planning (and thus promotes model-based control) whereas such planning is too costly for choices that are farther away from the end-point of a sequence of actions, hence the reliance on model-free values in that case. This is similar to what expert chess players report — in the middle of a game, they rarely plan many steps ahead. Instead, they rely on well-honed knowledge of what board positions are more favorable in the long run (for example, it is in general advantageous to have control of the center of the board). Close to the end of the game, however, all relevant sequences of actions by the opponent are considered through mental planning.

Incidentally, considering the game of chess as a model for decision making towards a long-term goal suggests a third decision-making system: in the beginning of a chess game, opening plays have been well documented, and are studied and memorized by experts, to use as needed. Indeed, Lengyel and Dayan (2008) suggested that this occurs in day-to-day decisions as well, where episodic memories of past successful actions are utilized for decision-making very early in learning, before either of the model-based or model-free systems have enough information with which to make a recommendation for the best action to perform.

Indeed, these conditions for goal-directed *versus* habitual control of animal behavior were what led Nathaniel Daw and colleagues to first suggest that goal-directed behavior can be mapped on to model-based reinforcement learning, whereas habitual behavior is a hallmark of model-free reinforcement learning. Moreover, to arbitrate between the two decision-making systems, they suggested that each system computes and reports not only values for different actions, but also its confidence in those values (standard error bars or confidence intervals on the estimated values). Arbitration between the systems can then utilize their relative confidence: at each point in time, the brain should trust the system that is most sure of its value estimate (Figure 2B).

One can interpret the behavioral findings in this light: the model-free system calculates its confidence as proportional to the amount of experience with this particular action in this situation (state). As in experimental science where larger sample sizes decrease the error in estimating the mean of a population, more experience with an action in a certain situation decreases the error in estimating the future rewards that can be expected if that action is chosen. The model-based system, in contrast, loses precision the farther into the future it has to simulate, given that the model it plans with is not fully accurate. The more iterations of using the model to figure out the future rewards contingent on an action (that is, its value), the more errors will accumulate in the estimate. Additionally, if the planning tree has many branches — that is, there are many possible future courses of action to evaluate and average over — computational noise or working memory limitations (or both) can inject noise into the value estimate. As such, the model-based system might put less confidence in estimates that required more computation, as a result of either the depth or the width of the simulated options.

This could explain why the action more distal from the reward outcome was deemed habitual, or model free, in Balleine and colleagues' action sequence study, even while the action proximal to reward was goal-directed/model-based. The more the animal needs to simulate forward, the more uncertain the model-based system, whereas the extensive training process in that experiment was sufficient to reduce the uncertainty of the model-free system for both actions (Figure 2B). According to the theory, arbitration between the two systems is done on a decision-by-decision basis: each system provides value estimates for each of the possible actions, complete with error bars on these estimates, and for each action, an arbitration system uses the value from the most confident system (or, alternatively, averages the two values inversely weighted by precision) to make the final action choice.

We note here that we have used the term 'uncertainty' not in the sense of 'risk', which is sometimes called uncertainty in the field of neuroeconomics, but rather in the sense of 'ambiguity'. Risk is a property of the stochasticity of the world — tossing a fair coin for a $1 reward if heads, and $1 loss if tails involves high risk as we are maximally uncertain about the outcome of the toss. However, this risk can be rationally and accurately averaged over to produce a mean expectation of $0. The uncertainty we are referring to is, instead, akin to ambiguity about the fairness of the coin, that is, not knowing the probability of reward itself. Reinforcement learning naturally averages over risk as it learns expected values (in the statistical notion of 'expectation'). But it is susceptible to ambiguity — the whole learning

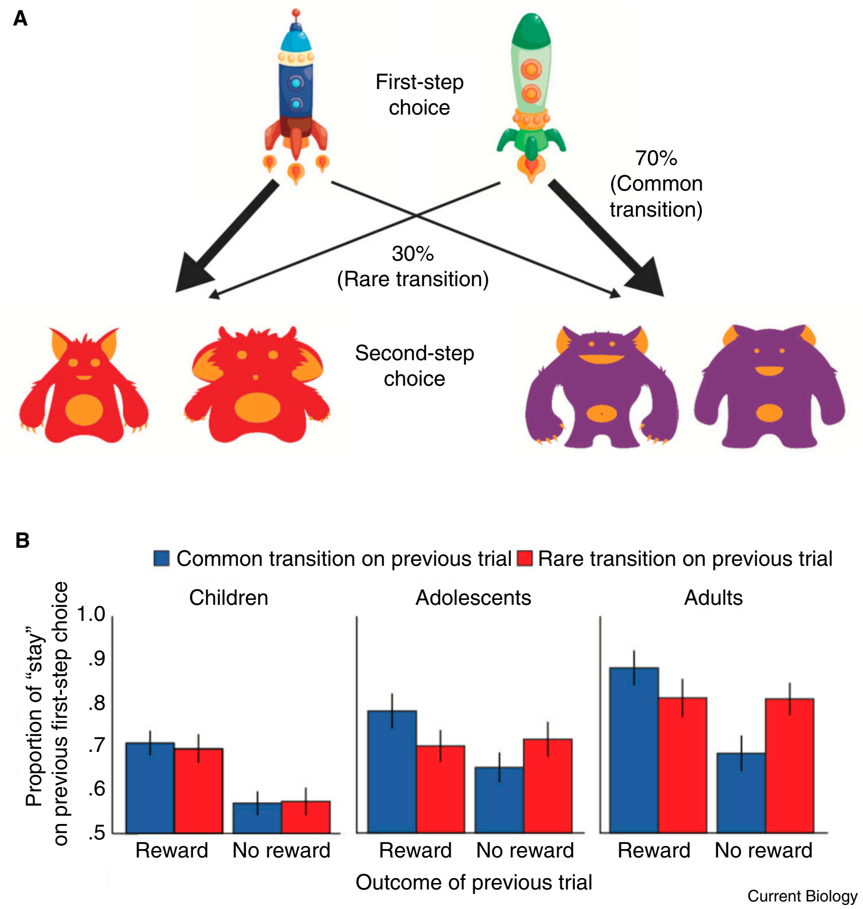


Figure 3. The 'two-step task'.
(A) In a version of the task used to study how model-based and model-free control changes over development, in the first stage, participants are presented with two options (spaceships) and asked to pick one. Each spaceship leads to one 'planet' with 70% chance and to the other with 30% chance. In the second stage, participants meet two 'aliens' on the planet they reached, and can choose which alien to ask for space treasure. Aliens are associated with different (and constantly changing) probabilities of having space-treasure reward to give. (B) The interaction between transition type and the effects of reward on subsequent choice that is characteristic of model-based learning emerges during adolescence and continues in adulthood. Figures adapted from Decker *et al*. (2016), reprinted from by Permission of SAGE Publications © 2016.

process is aimed at reducing ambiguity and estimating the expectation more and more accurately. As such, the hypothesis above, regarding each system estimating its own uncertainty, is different from ideas about reinforcement learning tracking risk and not only expected values.

Two distinct but interacting neural systems

Supporting the behavioral results described above, neuroscientific studies have shown that the brain harbors two largely parallel decision-making circuits. Habitual control using model-free reinforcement learning is associated with a decision-making loop from cortical afferents, through the dorsolateral striatum (putamen in primates), the ventrolateral thalamus, and back to cortex. Learning in cortico-striatal synapses in this loop is modulated by dopaminergic signals that convey prediction errors, as predicted by the theory. A parallel loop including the prelimbic cortex (dmPFC in primates), the dorsomedial striatum (caudate in primates), ventromedial thalamus and back to cortex is thought to mediate goal-directed control and model-based computations. Learning in the latter loop seems to be dopamine-independent (though recent evidence suggests that this may be an oversimplified view), but it relies on inputs from the hippocampus. Lesioning or inactivating either of these two loops will revert behavioral control to the remaining, intact system (Figure 1).

Although we have been discussing model-based and model-free reinforcement learning as if they are exclusive, the chess analogy highlights that the two systems might better work in tandem. For instance, one can use model-based planning to search several steps forward in a simulated tree of future possibilities, computing action values based on cached model-free values of the end-points of the search rather than simulating the game until its end. Similarly, early pruning of some branches of the search tree can be determined based on cached values. This may result in failure to exploit sequences of actions that initially lead to negative outcomes but eventually to winning the game (for example, sacrificing the queen for a later check mate). However, pruning can allow the decision maker to focus cognitive resources on simulating action sequences that are more likely to be advantageous in the long run. Of course, as the model-free system improves its estimates, pruning will be even more optimal.

In sum, although the distinction between two competing systems may seem simplistic, evidence for more mixed modes of decision making does not necessarily mean that the two systems do not exist as separate algorithms. It does suggest that competition and collaboration can take place at the level of individual actions and sweeps of forward planning in the model. Indeed, the two cortico-basal-ganglia loops are actually interconnected. This connection may underlie both the competition and cooperation between the two systems — the locus of arbitration between model-free and model-based decisions is not yet known. Additionally, the neurotransmitter norepinephrine has been associated with signaling environmental change (known as 'unexpected uncertainty'), which should trigger a shift from habitual, model-free behavior back to deliberative, goal-directed model-based reasoning. High levels of norepinephrine may therefore modulate the arbitration between the two systems, in addition to increasing learning rates for both the world model and the cached values.

A paradigm for studying model-based *versus* model-free decision making in humans

To test for model-based or model-free behavior in animals, the studies described above used a single change in the environment (most often: a change in the value of the outcome), and a subsequent test. However, the effects of planning in a model *versus* prediction-error based stimulus–response learning can be observed even without such abrupt changes in the environment. In particular, Nathaniel Daw has taken advantage of the specific algorithmic assumptions of model-based *versus* model-free learning to develop a task that can quantify, on a continuous scale, the extent to which human behavior conforms with either mechanism.

This paradigm is known as the 'two-step task' (Figure 3). On each trial, an initial choice between two options leads probabilistically to either of two second-stage states ('planets' in Figure 3, from an adaptation of the original task for use with children). At the second step, another choice between two options ('aliens') is made. Rewards at the second step are probabilistic and change slowly and independently for each of the four aliens. As a result, to maximize reward, participants must constantly learn which alien is best in each planet, and which planet is best overall. The transition probabilities from first to second step are fixed (and sometimes pre-trained): each first-level option leads to one planet 70% of the time ('common' transition), and to the other planet 30% of the time ('rare' transition).

Teasing apart model-based and model-free choice strategies occurs by observing how rewards received after a rare transition impact first-step choices on subsequent trials. In particular, after a reward is obtained in the second step, model-free reinforcement learning suggests that the first-step choice for that trial should be repeated in the subsequent trial, as this action was accompanied by a positive prediction error at the end of the two-step chain of actions. In contrast, model-based reinforcement learning updates the probability of reward for that alien, and by taking into account the transition structure of the task, knows that the best way to reach this now-more-valuable planet is by choosing the opposite first-step action. Therefore, model-free reinforcement learning predicts that choice in the first step should, in general, be consistent with feedback on the previous trial: repeat actions that were rewarded, avoid those that were not. Model-based reinforcement learning, in contrast, takes into account whether the outcome followed a rare transition or a common one to determine whether to repeat or avoid the previous action.

The two-step task has been leveraged to test the effects of a host of manipulations on the trade-off between model-based and model-free decision-making. For instance, Ross Otto and colleagues showed that a cold-presser stress manipulation decreases model-based decisions in the task, in which subjects with better working memory demonstrated more resilience in the face of stress. In a separate study, they showed that working memory load due to a dual task decreased model-based decision making, and that this competition between the two systems could be controlled on a trial-by-trial basis by imposing working memory constraints that restrict the availability of cognitive resources. These results can be explained by assuming that limiting computational resources, by stress or a concurrent task, renders the model-based system more uncertain in its (less rigorous and more noisy) calculations, and therefore control reverts to the model-free system. In a developmental study, Catherine Hartley and colleagues have shown that young children show dominance of model-free learning, with model-based strategies emerging in adolescence and increasing through adulthood (Figure 3B). Rodent variants of the task have also been developed, and are being used to dissect the neural mechanisms of both methods of decision-making, capitalizing on the ability to measure both model-based and model-free influences throughout many sessions of data collection.

Conclusion

Returning to the image of two competing decision-making systems sitting on our shoulders, it is now clear why neither is the devil. Values estimated by either model-based or model-free reinforcement learning are equally rational — though not necessarily equal. Indeed, because these systems use different algorithms to compute action values, they are differently susceptible to situational elements that can make decisions suboptimal or seemingly irrational. This may be why the brain employs both algorithms: when possible, it uses forward planning in a sophisticated, prospective, model-based system. However, the high cognitive demands of such decision making require exclusive use of precious resources, preventing other concurrent tasks. Luckily, our good-old habitual, model-free reinforcement learning system is ever ready to jump in, and to allow us to maintain the illusion that we can multi-task.

FURTHER READING

Balleine, B.W., Garner, C., Gonzalez, F., and Dickinson, A. (1995). Motivational control of heterogeneous instrumental chains. J. Exp. Psych. Anim. Behav. Proc. *21*, 203–217.

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

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

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

Dickinson, A., and Balleine, B. (1994). Motivational control of goal-directed action. Anim. Learn. Behav. *22*, 1–18.

Dickinson, A., Balleine, B., Watt, A., Gonzalez, F., and Boakes, R.A. (1995). Motivational control after extended instrumental training. Anim. Learn. Behav. *23*, 197–206.

Killcross, S., and Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. Cereb. Cortex *13*, 400–408.

Lengyel, M., and Dayan, P. (2008). Hippocampal contributions to control: the third way. In Proceedings of the 21st International Conference on Neural Information Processing Systems (Red Hook, NY: Curran Associates Inc.) pp. 889–896.

Otto, A.R., Gershman, S.J., Markman, A.B., and Daw, N.D. (2013). The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. Psych. Sci. *24*, 751–761.

Otto, A.R., Raio, C.M., Chiang, A., Phelps, E.A., and Daw, N.D. (2013). Working-memory capacity protects model-based learning from stress. Proc. Nat. Acad. Sci. USA *110*, 20941–20946.

Yin, H.H., Ostlund, S.B., Knowlton, B.J., and Balleine, B.W. (2005). The role of the dorsomedial striatum in instrumental conditioning. Eur. J. Neurosci. *22*, 513–523.

Princeton Neuroscience Institute, Washington Road, Princeton, NJ 08544, USA.
E-mail: nicoled@princeton.edu (N.D.); yael@princeton.edu (Y.N.).