COMMENTARY

Of goals and habits

Nathaniel D. Daw[1]

Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ 08544

In the television series *The Wire*, addicts Bubbles and Johnny regularly engage in bizarre and elaborate schemes to obtain drugs, ranging from robbing an ambulance to stealing drugs by lowering a fishing line from a rooftop. That these fictional crimes can be both so meticulously planned and yet focused on such narrow, shortsighted goals highlights a gap in our understanding of how the real brain deploys deliberative vs. automatic mechanisms to make decisions. On a standard account, people can deliberatively evaluate the consequences of candidate actions, which gives us our flexibility to dream up novel plans. Alternatively, the brain can crystallize repeatedly successful behaviors into habits: learned reflexes that free up resources by executing the behaviors automatically (although at the expense of inflexibility and, it is believed, underpinning pathological compulsions). As with most dichotomies, the problem with this view is that the world is not so black and white. Much as the drug-seeking behavior of addicts seems not to fit into either category, for healthy behaviors also, neither of these two sorts of decision making is very practical on its own. In PNAS, Cushman and Morris suggest a hybrid of these mechanisms, and show behavioral evidence that humans use it to plan actions (1).

The study of Cushman and Morris (1) draws on recent advances using computational models of learning to make these strategies explicit enough that their hallmarks can be measured in choice behavior. Decisions are often modeled as determined by one's estimates of the rewards expected from different options. There are many different methods to compute these decision variables. Deliberative planning can be formalized by an algorithmic family, called "model-based reinforcement learning" (2), which evaluates candidate sequences of actions much like a chess computer does, by exhaustively searching the "tree" of their future consequences, generated using a learned model of the task contingencies (like the rules of chess or the map of a maze) (Fig. 1*A*).

The key feature of habits, in this view, is that they instead rely on a simpler summary of the end results of this computation, such as the overall long-run reward expected following some action (2). This precomputation gives them both their simplicity and inflexibility. These summaries do not actually need to be derived from exhaustive computations using a model, but instead can be learned directly—although slowly—"model-free" from experience (3).

Cushman and Morris (1) propose a hybrid of these strategies, in which each system simplifies the problem faced by the other (Fig. 1*B*). The hypothesis is that, in a multistep decision problem, model-free learning selects a goal or subgoal, then model-based planning figures out how to get there. If I wanted to travel to Paris, a goal might be the airport; in chess, it might be forking the opponent's queen. This is a useful division of labor because, in general, finding the best action using a model-based search is too complicated because of the many possible trajectories of future action. However, given a single goal, finding a good path to it can be more manageable. Meanwhile, although model-free learning is computationally simple, it learns slowly in large, multistep tasks. Narrowing its attention to a smaller, more abstracted problem—choice between a few candidate goals—lets its simplicity shine.

The key to the experiments of Cushman and Morris (1) is that these strategies learn in different ways from experience in trial-and-error decision tasks (4). Indeed, just how behavior changes following a single, carefully arranged trial of experience can reveal things like whether a goal was updated, and if so how. The authors (1) harnessed the efficiency of online data collection to zero-in on these rare informative choices across a large number of subjects.

In the studies (1), participants chose between actions, which led to intermediate situations represented by different colors, and then to monetary reward. If a choice that led to the blue intermediate goal was followed by a large pay-off, subjects tended to choose actions leading to blue again later. This was true even if the later action needed to get to blue was a different or even completely novel one, which suggests that planning to get to blue was model-based. This is because model-based planning, by evaluating actions in terms of their predicted consequences, can generalize across different ways to achieve the same result (5). If I want to get to the airport, model-based learning can find a route there from anywhere, whereas model-free learning can only repeat previously successful routes.

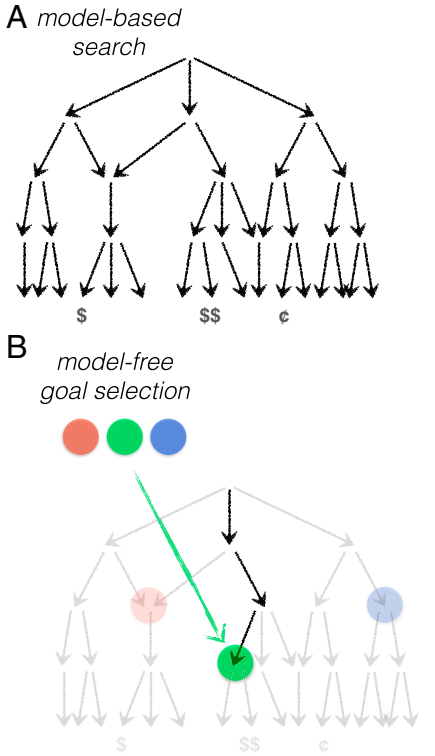

Fig. 1. (*A*) A stylized choice problem with a large tree of possible future action sequences, posing a difficult search problem for model-based planning. (*B*) Cushman and Morris' (1) hypothesis: model-based planning can be simplified by initial model-free selection which goal (here, color) to seek.

Surprisingly, a rewarded choice for blue was repeated even if the subjects in the study (1) had failed to achieve the intended blue goal on the initial choice, so that this reward was irrelevant to the true value of the goal. This apparent mistake is exactly what model-free selection predicts, because it assesses the success of a strategy (here, choosing blue) in terms of the reward received. Four experiments

Author contributions: N.D.D. wrote the paper.

The author declares no conflict of interest.

See companion article 10.1073/pnas.1506367112.

[1]Email: ndaw@princeton.edu.

replicated this pattern and controlled for several alternative interpretations.

This research relates to a broader literature on the tricks the brain uses to simplify multistep choice (6). Cushman and Morris' (1) hypothesis is an instance of a set of approaches called hierarchical reinforcement learning, which decompose a multistep decision problem into a nested set of choices at different levels of temporal abstraction. There is mounting neural and behavioral evidence that the brain uses such decomposition (7, 8). However, in principle, such abstraction is indifferent to the learning mechanism: decisions at any level could be approached by model-based or model-free learning (or any combination), with this particular division being just one option. Indeed, a previous study investigating such abstraction was used to suggest that the top-level choices (over goals, in the present terminology) were actually model-based rather than, as Cushman and Morris suggest, model-free (8). Both might be right: the current experiments (1) do not rule out the coexistence of other approaches. However, the earlier experiment omitted key controls included in the new study, so those results are also compatible with model-free goal selection.

A similar point applies to the lower level of Cushman and Morris' (1) model. Several other researchers have envisioned analogous lower-level choices as essentially model-free: "chunked," stereotyped behavioral sequences, like a tennis serve or a dance move. Some evidence suggests that this sort of memorization of extended routines may be the real foundation of habits (9, 10). However, such stereotyped sequences stand in contrast to Cushman and Morris' (1) flexible model-based goal-seeking to find a desired result. For example, in one of their experiments, subjects reach the blue goal by finding groups of numbers that add up to 21. Is temporal abstraction particularly helpful for learning new tasks, generalization, and transfer (as Cushman and Morris' model would be), or is it mainly about streamlining the control of well-learned behaviors (as with chunking)? Are these two separate mechanisms? Another unresolved issue is how subgoals themselves are discovered. The efficiencies of hierarchical control are only realized given an appropriate (and limited) set of subgoals: the airport is a broadly useful subgoal but many other places are not (11).

Another area in which the traditional dichotomy between model-based and model-free control has raised many questions is the systems' neural substrates. Although brain lesion studies show a dissociation between areas associated with either sort of control (12), neuroimaging and unit recording studies in intact brains tend not to find such strict separation (13). Instead, neural correlates of decision variables in many parts of the brain show a mixture of both types of responses (4, 14). Hybrid schemes, like that of Cushman and Morris (1), in which the two mechanisms interact toward solving the same problem, may help to rationalize these results. It will be particularly interesting to study how these neural signals behave in the new decision tasks, which exercise this interaction.

The study of Cushman and Morris draws on recent advances using computational models of learning to make these strategies explicit enough that their hallmarks can be measured in choice behavior.

Returning to the hapless Bubbles and Johnny, this research may point toward the resolution of a longstanding problem in the study of drug abuse and other disorders of compulsion. It is often argued that the compulsive nature of drugs might be rooted in the inflexible character of habits (15). Indeed, patients with a range of compulsive disorders show abnormal dominance of model-free over model-based learning on a task similar to that of Cushman and Morris (16), and the neuromodulator dopamine (a common target of addictive drugs) is believed to drive model-free learning (17). However, although this might explain how addicts acquire some stereotyped actions involved in drug consumption, like raising a cigarette to one's mouth, simply repeating previously rewarded actions wholly fails to explain many more deliberative and goal-oriented drug-seeking behaviors (18), as dramatized by Bubbles' bizarre (and presumably wholly novel) fishing escapade. Equally, on the traditional, dichotomous account, model-based planning should have no interest in attaining a poisonous substance that happens to affect preferences in a separate model-free system. This paradox is resolved if that same model-free learning mechanism extends to choosing goals for the model-based system. Other previously disjoint phenomena of drug abuse, such as craving and attentional biases, may also be seen as further manifestations of these abstract habits.

Finally, beyond compulsion, researchers have found signs of dual-system decision making in many diverse situations where we seem to be of two minds, from moral dilemmas to procrastination to racism. The interactions between our more deliberative and more automatic impulses, of the sort documented by Cushman and Morris (1), may therefore ultimately shed light on many of the most puzzling of human experiences.

1 Cushman F, Morris A (2015) Habitual control of goal selection in humans. *Proc Natl Acad Sci USA*, 10.1073/pnas.1506367112.

2 Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* 8(12):1704–1711.

3 Sutton RS (1988) Learning to predict by the methods of temporal differences. *Mach Learn* 3(1):9–44.

4 Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69(6):1204–1215.

5 Doll BB, Duncan KD, Simon DA, Shohamy D, Daw ND (2015) Model-based choices involve prospective neural activity. *Nat Neurosci* 18(5): 767–772.

6 Huys QJ, et al. (2015) Interplay of approximate planning strategies. *Proc Natl Acad Sci USA* 112(10):3098–3103.

7 Botvinick MM, Niv Y, Barto AC (2009) Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition* 113(3):262–280.

8 Dezfouli A, Balleine BW (2013) Actions, action sequences and habits: Evidence that goal-directed and habitual action control are hierarchically organized. *PLOS Comput Biol* 9(12):e1003364.

9 Dezfouli A, Balleine BW (2012) Habits, action sequences and reinforcement learning. *Eur J Neurosci* 35(7):1036–1051.

10 Smith KS, Graybiel AM (2013) A dual operator view of habitual behavior reflecting cortical and striatal dynamics. *Neuron* 79(2):361–374.

11 Solway A, et al. (2014) Optimal behavioral hierarchy. *PLOS Comput Biol* 10(8):e1003779.

12 Balleine BW, O'Doherty JP (2010) Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* 35(1):48–69.

13 Doll BB, Simon DA, Daw ND (2012) The ubiquity of model-based reinforcement learning. *Curr Opin Neurobiol* 22(6):1075–1081.

14 Bromberg-Martin ES, Matsumoto M, Hong S, Hikosaka O (2010) A pallidus-habenula-dopamine pathway signals inferred stimulus values. *J Neurophysiol* 104(2):1068–1076.

15 Everitt BJ, Robbins TW (2005) Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nat Neurosci* 8(11): 1481–1489.

16 Voon V, et al. (2015) Disorders of compulsivity: A common bias towards learning habits. *Mol Psychiatry* 20(3):345–352.

17 Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* 16(5):1936–1947.

18 Tiffany ST (1990) A cognitive model of drug urges and drug-use behavior: Role of automatic and nonautomatic processes. *Psychol Rev* 97(2):147–168.