**Habitual control of goal selection**

Fiery Cushman1 and Adam Morris2

1*Department of Psychology, Harvard University*

2*Department of Cognitive, Linguistic and Psychological Science, Brown University*

**The distinction between habitual and goal-directed action is fundamental to behavioral research1-3. Habits form as stimulus-response pairings are “stamped in” following reward. They enable computationally efficient decision making, but at the cost of behavioral flexibility. In contrast, goal-directed behavior requires planning over a causal model. This enables more flexible decision-making, but at a potentially severe computational cost. Exhaustive search over candidate goals becomes prohibitive as the complexity of the model grows. Thus, a key requirement for goal-directed action is to efficiently select candidate goals with a high likelihood of reward. Here, we provide evidence for a potential solution implemented by humans: Habitual control over the process of goal selection. Although many existing treatments of the distinction between habitual and goal-directed action emphasize their competition over behavioral control4, our results illustrate a codependence between the systems5. The role of habitual control in goal-directed action explains diverse phenomena including the automatic selection of goals under contextual cuing6,7 and the basis of practice effects in cognitive skills**8-10**.**

Our approach depends on a formalization of habitual and goal-directed behavior derived from the reinforcement learning (RL) framework11. Model-based RL maintains an explicit causal model of the world and uses it to choose actions by assessing their likely consequences. In contrast, model-free RL does not maintain an explicit causal model, and thus does not allow planning. Rather, like Thorndike’s law of effect3, it assigns value to candidate actions based on their context-dependent history of reward. The resulting stimulus-response habits are globally adaptive, but may exhibit local irrationality12,13.

RL models capture several core features of learning and choice in humans1,12,14. Elements of model-free RL, including prediction-error updating and temporal difference learning, are implemented in midbrain dopamine system15-17. Human behavior also relies extensively on model-based planning towards goals18, often arranged hierarchically (plug in the machine to grind the beans to make the coffee, etc.)19. Once selected, goals can dramatically reduce the computation necessary for model-based action selection20,21. The resulting policies are then available for reuse and recombination, further reducing computational demands22. First, however, an appropriate goal must be selected. Full model-based evaluation of candidate goals often imposes a prohibitive computational cost. In principle, given the superordinate goal of making coffee, an infinite number of subordinate goals might be entertained and evaluated. How do we efficiently alight upon the idea of grinding beans?

One potential solution is to allow model-free control over goal selection. In other words, the goal of grinding beans might be “stamped in” due to the history of reward associated with this goal in past coffee-making episodes. Subsequent planning to achieve the selected goal could then proceed in a model-based fashion, or by other methods. This enables computationally tractable goal selection while maintaining the possibility of flexible planning towards the goal.

Colloquially, this proposal captures the notion of a “habit of thought”: Model-free control can contribute to the effective deployment of cognitive routines which themselves transcend learned stimulus-response pairings22. Consistent with this proposal, recent research emphasizes the pervasive role of model-free control in related elements of higher-level cognition23,24, including the gating of working memory25 and the construction of hierarchical task representations26. Such proposals offer an appealing functional explanation for the neuronal connections between striatum and frontal cortex27. The possibility of habitual control over goal selection also accords with several formal models of RL28 and human cognition2,29.

In order to test this possibility we adapted a multistep choice paradigm from prior research12. The original paradigm behaviorally dissociates the influence of habitual (model-free) and goal-directed (model-based) control on choice, and is well-validated4,30-32. Our modification allows us to index not only model-free value assignment to actions (as in the original task) but also model-free value assignment to goals, which may be subsequently pursued via model-based planning (Figure 1a). At Stage 1 of each trial participants make a choice between two options drawn from the set (1, 2, 3, 4). These choices stochastically transition to a second set of three states (blue, red, green). Finally, Stage 2 states deterministically transition to three unique reward distributions. The rewards change gradually over the course of the experiment. Thus, participants are motivated to choose Stage 1 options that maximize the likelihood of transitioning to the current reward-maximizing final state. For task details, see Supplementary Materials.



**Figure 1: Task structure. a**, In Experiment 1 participants performed a two-stage Markov decision task. **c**, In Experiment 2, Stage 2 options varied along two orthogonal dimensions: shape and color. On each trial the participant was cued whether rewards would be determined by shape or color for that trial. **b**, Experiment 3 was identical to Experiment 1 except that participants were pre-trained on a set of transitions from Stage 0 to Stage 1, and then presented with Stage 0 choices during the main task only on critical “congruent goal” test trials.

Our analysis depends on a critical subset of trials. For example, a participant is presented with the choice set (1, 2) at Stage 1 and chooses 1. Because 1 typically leads to the blue state, we assume that this participant’s goal was to transition to blue. On our critical trials, however, they experience a low-probability transition to the green state, and then experience a very large reward. A model-based system would discard this information because transitions to the green state are equally likely from all Stage 1 options. This renders forward planning toward green irrelevant. In contrast, model-free value update would increase the likelihood of selecting 1 on subsequent trials due to the positive reward history12. Our interest, however, is in the model-free assignment of value to a goal; in this case, the goal of transitioning to blue. If the experience of reward increases the likelihood of selecting blue as a goal, then participants should exhibit a greater likelihood of choosing 3 on the subsequent trial (when paired with either 2 or 4). Conversely, the experience of punishment should decrease the likelihood of choosing 3. This influence of the reinforcement history of choosing 1 on the subsequent choice of 3 cannot be explained by model-free update of a value to the specific action (choosing 1); rather, it depends on the assignment of value to their shared goal (getting to blue).

We assessed trials of this type by comparing instances when the participant experienced reward vs. punishment following low-probability transition to the green state. The mean proportion of trials on which participants selected the congruent-goal option following positive reward (89%) was significantly greater than following negative reward (69%) *t*(134)=-12.5, *p*<.0001 (Figure 2a). Additional supporting analyses for all experiments are presented in supplementary materials.



**Figure 2: Results.** Bars represent that proportion of trials on which participants chose the goal-congruent action, averaged across participants. Whiskers indicate the standard error of the mean of these proportions across participants. **a-c** show results from Experiments 1-3, respectively. **b**, same trial type bars show the effect of reward experienced on trials of the same type (shape vs. color) while different trial type bars show the effect of reward experienced on trials of the opposite type.

The evidence from Experiment 1 is ambiguous between two interpretations. It may be that people assign value to the selection of a goal (e.g., “choose blue”), or it may be that people assign value directly to the congruent-goal Stage 1 option (e.g., “choose option 3”). Experiment 2 was designed to disambiguate these possibilities (Figure 1b). Specifically, Stage 2 states were arranged in a 2×2 design crossing color (red versus blue) and shape (circle versus square), with a fifth state that differed on both dimensions (a green triangle). Each trial was defined as a “color trial” or “shape trial”, with the trial type dictating the deterministic transitions to drifting rewards, and participants were cued to trial type at the beginning of each trial. Thus, three reward distributions were accessible on color trials, while three independent reward distributions were accessible on shape trials. If model free value is assigned to goals, it should only influence choice on subsequent trials of the same type (i.e., “color trial” vs. “shape trial”). Consistent with this prediction, we replicated our result from Experiment 1 for same-type trials: After a low probability transition to the green triangle state, participants were more likely to choose the congruent goal Stage 1 option on a subsequent same-type trial following reward (83%) compared with punishment (76%) *t*(302)=-4.82, *p*<.001. On different-type trials, however, there was no significant difference (positive: 50%; negative: 47%) *t*(282)=-.94, *p=.35* (Figure 2b).

The results of Experiments 1 and 2 are consistent with model-free control over goal selection in which the habitually selected goal then participates in a process of forward planning over a causal model of the transition structure of the task. They are also consistent, however, with some hierarchical reinforcement learning models that do not invoke true model-based planning22,33. These models assume that “goal states” establish internally represented contexts that bias model-free stimulus response associations. We designed Experiment 3 to test whether the goals selected in our paradigm could be flexibly integrated with knowledge of independent state transitions, a hallmark of true planning.

Participants were first trained on a deterministic set of transitions between four Stage 0 options (A, B, C, D) and the same four Stage 1 options used in Experiment 1 (Figure 1c). During this training phase the Stage 1 comprised the terminal states. Then, participants trained on and performed the same task used in Experiment 1, without any involvement of Stage 0 choices. Finally, we tested each participant in a set of critical trials—those involving a low-probability transition followed by a goal-congruent choice—but presented participants with a pair of Stage 0 choices in place of the ordinary Stage 1 choices. In order to integrate information about a desired goal with the set of Stage 0 choices, participants were therefore required to engage in forward planning over the learned transition structure between Stage 0 and Stage 1. Here, again, we found that participants were significantly more likely to choose the congruent-goal option following positive reward (85%) than following negative reward (69%) *t*(172)=-9.17, *p*<.0001 (Figure 2c).

Experiment 3 undermines another alternative interpretation of our results. Possibly, the statistical structure of state transitions could support associative pairings between Stage 1 options34. For instance, participants might associate options 1 and 2 because they share a common high-probability transition to the blue state in Stage 2. Experiment 2 shows that any such associative model must account for the higher-level contextual state of trial type. Experiment 3 provides the strongest evidence against this alternative, however, because there is no basis for statistical association between Stage 0 choices.

In sum, we find that goal selection is partially determined by model-free value representations derived from reward history. These goals are subsequently used during model-based planning according to an internally represented causal model of the task structure. Thus, while our proposal relies upon the conceptual distinction between habitual (model-free) and goal-directed (model-based) behavioral control, it also demonstrations an area of mutual dependence between these mechanisms.

This proposed integration captures several phenomena that blend canonical features of habits and goals. Many studies find that contextual cues can trigger goal pursuit outside of conscious awareness6, consistent with the operation of stimulus-response habits in the process of goal selection. In cases of “utilization behavior” among individuals with insult to prefrontal cortex, goal-directed behavior may be intrusive or inappropriately invoked based on contextual cues7. Among neurotypical individuals, “functional fixedness” describes the tendency to consider a limited set of candidate means-end relationships based on past experience with a tool10. Finally, it is commonly observed in educational settings that the execution of controlled cognitive processes improves with practice—in other words, that learning complex tasks requires the incremental acquisition of appropriate habits of thought8,9. It is widely recognized that humans’ representations of complex tasks are organized hierarchically into goals and subgoals2,19,22. Proficiency may often depend upon the gradual acquisition of habitual subgoal selection given the contextual state of a superordinate goal.

References

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

2 Norman, D. A. & Shallice, T. *Attention to action*. (Springer, 1986).

3 Thorndike, E. L. Animal intelligence: An experimental study of the associative processes in animals. *Psychological Monographs: General and Applied* **2**, i-109 (1898).

4 Otto, A. R., Gershman, S. J., Markman, A. B. & Daw, N. D. The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychol Sci* **24**, 751-761, doi:10.1177/0956797612463080 (2013).

5 Gershman, S. J., Markman, A. B. & Otto, A. R. Retrospective Revaluation in Sequential Decision Making: A Tale of Two Systems. (2012).

6 Huang, J. Y. & Bargh, J. A. The Selfish Goal: Autonomously operating motivational structures as the proximate cause of human judgment and behavior. *Behavioral and Brain Sciences* **37**, 121-135 (2014).

7 Lhermitte, F. ‘Utilization behaviour’and its relation to lesions of the frontal lobes. *Brain* **106**, 237-255 (1983).

8 Perkins, D. N. & Salomon, G. Are cognitive skills context-bound? *Educational researcher* **18**, 16-25 (1989).

9 Sfard, A. On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educational studies in mathematics* **22**, 1-36 (1991).

10 Adamson, R. E. Functional fixedness as related to problem solving: A repetition of three experiments. *Journal of experimental psychology* **44**, 288 (1952).

11 Sutton, R. S. & Barto, A. G. Reinforcement learning. *Journal of Cognitive Neuroscience* **11**, 126-134 (1999).

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

13 Dickinson, A., Balleine, B., Watt, A., Gonzalez, F. & Boakes, R. A. Motivational control after extended instrumental training. *Learning & behavior* **23**, 197-206 (1995).

14 Simon, D. A. & Daw, N. D. Neural correlates of forward planning in a spatial decision task in humans. *The Journal of neuroscience* **31**, 5526-5539 (2011).

15 McClure, S. M., Berns, G. S. & Montague, P. R. Temporal prediction errors in a passive learning task activate human striatum. *Neuron* **38**, 339-346 (2003).

16 O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H. & Dolan, R. J. Temporal difference models and reward-related learning in the human brain. *Neuron* **38**, 329-337 (2003).

17 Bayer, H. M. & Glimcher, P. W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129-141 (2005).

18 Botvinick, M. M. Hierarchical models of behavior and prefrontal function. *Trends in cognitive sciences* **12**, 201-208 (2008).

19 Lashley, K. S. *The problem of serial order in behavior*. 112-131 (Wiley, 1951).

20 Boyan, J. A. & Moore, A. W. in *ICML.* 63-70.

21 Zhang, N. L. & Zhang, W. in *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence.* 489-494 (Morgan Kaufmann Publishers Inc.).

22 Botvinick, M. M., Niv, Y. & Barto, A. C. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition* **113**, 262-280 (2009).

23 Dayan, P. How to set the switches on this thing. *Current Opinion in Neurobiology* (2012).

24 Graybiel, A. M. Habits, rituals, and the evaluative brain. *Annu. Rev. Neurosci.* **31**, 359-387 (2008).

25 O'Reilly, R. C. & Frank, M. J. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation* **18**, 283-328 (2006).

26 Badre, D. & Frank, M. J. Mechanisms of Hierarchical Reinforcement Learning in Cortico–Striatal Circuits 2: Evidence from fMRI. *Cerebral Cortex* **22**, 527-536 (2012).

27 Miller, E. K. The prefrontal cortex and cognitive control. *Nature Reviews* **1**, 59-65 (2000).

28 Sutton, R. S., Precup, D. & Singh, S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* **112**, 181-211 (1999).

29 Cooper, R. & Shallice, T. Contention scheduling and the control of routine activities. *Cognitive Neuropsychology* **17**, 297-338 (2000).

30 Otto, A. R., Skatova, A., Madlon-Kay, S. & Daw, N. D. Cognitive Control Predicts Use of Model-based Reinforcement Learning. (2014).

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

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

33 Ribas-Fernandes, J. Ú. J. F. *et al.* A neural signature of hierarchical reinforcement learning. *Neuron* **71**, 370-379 (2011).

34 Daw, N. D. & Shohamy, D. The cognitive neuroscience of motivation and learning. *Social Cognition* **26**, 593-620 (2008).