Title: Habitual Control of Goal Selection in Humans

**Authors:** F. Cushman1\* and A. Morris2

**Affiliations:**

1 Department of Psychology, Harvard University.

2 Department of Cognitive, Linguistic and Psychological Sciences, Brown University.

\*Correspondence to: cushman@fas.harvard.edu

**Abstract**: Goal-oriented planning is a hallmark of human behavior. A central challenge is to efficiently select goals from the infinite space of potential candidates. We describe a solution grounded in computational models of reinforcement learning: Habitual control over the process of goal selection. This approach exploits the computational efficiency of habits to select a goal, while preserving the flexibility of planning processes once a goal has been selected. We find experimental evidence that human participants spontaneously employ this solution. While many existing treatments of the distinction between habitual and goal-directed action emphasize their competition over behavioral control, our results illustrate a codependence between the systems in guiding human action.

**Main Text:** The distinction between habitual and goal-directed action is fundamental to behavioral research(*1-4*). It can be formalized within reinforcement learning (RL) framework (*5*), which captures important features of learning and choice in humans(*1,6,7*). Habits form as stimulus-response pairings are “stamped in” following reward (*2*). They enable computationally efficient decision making, but at the cost of behavioral flexibility. Habitual control corresponds to model-free RL methods, which assign value to candidate actions based on their context-dependent history of reward. In contrast, model-based RL methods involve planning over a causal model linking candidate actions to their expected outcomes. Model-based methods can therefore support goal-directed planning. This enables more flexible decision-making, but the associated computational costs can become severe as the complexity of the model grows.

Elements of model-free RL, including prediction-error value update and temporal difference learning, are implemented in the midbrain dopamine system (*8-10*). Human behavior also relies extensively on model-based planning towards goals, often arranged hierarchically (plug in the machine to grind the beans to make the coffee, etc.) (*4,11,12*). Once selected, goals can dramatically reduce the computation necessary for model-based action selection (­*13,14*). The resulting policies are then available for reuse and recombination, further reducing computational demands (*15*). 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 next relevant goal: ground beans?

One potential solution is to allow model-free control over goal selection. In other words, the goal of ground 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 potential for flexible planning towards the selected goal. Colloquially, this proposal captures the notion of a “habit of thought”: Model-free control can contribute to the effective deployment of model-based cognitive routines that ultimately transcend learned stimulus-response pairings. Consistent with this proposal, recent research emphasizes the pervasive role of model-free control in related elements of higher-level cognition(*16-17*), including the gating of working memory (*18*) and the construction of hierarchical task representations(*19*). Such proposals offer an appealing functional explanation for the neuronal connections between striatum and frontal cortex(*20*). The possibility of habitual control over goal selection also accords with formal approaches to the RL(*21*) and human cognition(*4*).

In order to test this possibility we adapted a multistep choice paradigm from prior research(*6*). The original paradigm behaviorally dissociates the influence of habitual (model-free) and goal-directed (model-based) control on choice, and is well-validated(*22-25*). 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 (Fig. 1A). At Stage 1 of each trial participants make a choice between two actions drawn from the set [1,2,3,4]. These choices stochastically transition to Stage 2 states from the set [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. Participants received with detailed instructions and practice trials, including information about the stochastic transitions between Stage 1 and Stage 2. For task details, see Supplementary Materials.



**Fig. 1.** (**A**) In Experiment 1 participants performed a two-stage Markov decision task. They were presented with two possible Stage 1 actions drawn from a set of four. These transitioned with variable probabilities to a set of Stage 2 actions, which then transitioned deterministically to a set of drifting reward distributions. (**B**) The logic of the experiment depends on a subset of trials. For instance, participants might be presented with the choice set (1,2) in a setup trial. Upon selecting action 1, they experience a low-probability transition to the green state followed by a large reward. A model-free influence on goal selection uniquely predicts an increase in the selection of action 3 on the subsequent critical trial, because actions 1 and 3 share the common goal state of blue.

Our analysis depends on a critical subset of trials (Fig. 1B). For example, a participant is presented with the choice set [1,2] at Stage 1 and chooses action 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 history10. 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 in a setup trial. The mean proportion of trials on which participants selected the congruent-goal action following positive reward (89%) was significantly greater than the proportion following negative reward (69%) *t*(134)=-12.5, *p*<.0001 (Fig. 2A). Additional supporting analyses for all experiments are presented in Supplementary Materials.



**Fig. 2.** Bars represent the proportion of trials on which participants chose the congruent-goal action, averaged across participants. Whiskers indicate the standard error of the mean of these proportions across participants. (**A-C**) Results from Experiments 1-3, respectively.

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 action (e.g., “choose option 3”). Experiment 2 was designed to disambiguate these possibilities (Fig. 3). Specifically, Stage 2 states were arranged in a 2(color: red vs. blue)×2(shape: circle vs. square) design, 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. 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. This made color goals relevant only to color trials, and shape goals relevant only to 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 action 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* (Fig. 2B).



**Fig. 3.** In Experiment 2, Stage 2 states varied along two orthogonal dimensions: shape and color. On each trial the participant was cued whether rewards would be determined by shape or by color. We predicted that rewards obtained following low-probability transitions to the green state would only influence subsequent choice on critical trials of the same trial type (shape vs. color). This is because the goals selected in each of two trials can only match when their trial types are identical.

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 task’s transition structure. They are also consistent, however, with some hierarchical reinforcement learning models that do not invoke true model-based planning (*26,27*). 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 (Fig. 4A). During this training phase the Stage 1 options 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 following the setup of a low-probability transition followed by a congruent-goal choice—but presented participants on critical trials with a pair of Stage 0 choices in place of the ordinary Stage 1 choices (Fig. 4b). In order to integrate information about a desired goal with the set of Stage 0 choices, participants were 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 action following positive reward (85%) than following negative reward (69%) *t*(172)=-9.17, *p*<.0001 (Fig. 2C). These results also speak against an alternative interpretation of Experiments 1 and 2 according to which statistical structure of state transitions could support associative pairings between congruent-goal Stage 1 options, and thus associative transfer of reward values(*7*). In Experiment 3 there is no such basis for statistical association between Stage 0 actions analyzed on the critical trial and the Stage 1 action rewarded on the setup trial.



**Fig. 4.** Experiment 3 was modeled on the design of Experiment 1, except that (**A**) participants performed a pre-training in which they learned deterministic transitions between Stage 0 and Stage 1 choices, and (**B**) on critical trials the Stage 0 choices were selectively reintroduced. Thus, in order to make successful choices on critical trials, participants were required to choose a Stage 0 option that would lead to their preferred Stage 1 state.

Our experiments indicate that goal selection is partially determined by model-free value representations derived from reward history. These goals are subsequently used during model-based planning over 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 demonstrates an area of mutual dependence between these mechanisms.

This proposed integration captures several phenomena that blend canonical features of habits and goals. Contextual cues can trigger goal pursuit outside of conscious awareness (*28*), 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 cues (*29*). Among neurotypical individuals, “functional fixedness” describes the tendency to consider a limited set of candidate means-end relationships based on past experience with a tool(*30*). Finally, it is 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 thought (*31*). It is widely recognized that humans’ representations of complex tasks are organized hierarchically into goals and subgoals (*4,11,12*). Task proficiency may depend partially upon the gradual acquisition of habitual subgoal selection given the contextual state of a superordinate goal.

References and Notes:

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

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

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

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

5. Sutton, R. S. & Barto, A. G. Introduction to reinforcement learning. *MIT Press*, (1998).

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

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

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

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

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

11. K. S. Lashley, *The problem of serial order in behavior*. (Wiley, New York, 1951).

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

13. Boyan, J. A. & Moore, A. W. Learning evaluation functions for large acyclic domains. *Proceedings of the International Conference on Machine Learning.* 63-70 (1996).

14. Zhang, N. L. & Zhang, W*.* Fast value iteration for goal-directed markov decision processes. *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence* 489-494 (1997).

15. Collins, A. G. & Frank, M. J. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review* **120**, 190 (2013).

16. Dayan, P. How to set the switches on this thing. *Current Opinion in Neurobiology* **22**, 1068-1074 (2012).

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

18. 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).

19. Collins, A. G. & Frank, M. J. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review* **120**, 190 (2013).

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

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

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

23. A. R. Otto, A. Skatova, S. Madlon-Kay, N. D. Daw, Cognitive Control Predicts Use of Model-based Reinforcement Learning. *Journal of Cognitive Neuroscience* **27**, 319-333(2015).

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

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

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

27. J. Ú. J. F. Ribas-Fernandes, A. Solway, C. Diuk, J. T. McGuire, A. G. Barto, Y. Niv, M. M. Botvinick, A neural signature of hierarchical reinforcement learning. *Neuron* **71**, 370-379 (2011).

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

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

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

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

**Acknowledgments:** The data reported in this paper are presented in Supplementary Materials. We thank Michael Frank, Samuel Gershman and Josiah Nunziato for their advice and assistance. This research was supported by grant N00014-14-1-0800 from the Office of Naval Research.

Supplementary Materials:

Materials and Methods

Figures S1-S#

Tables S1-S#

Movies S1-S#

Audio Files S1-S#

External Databases S1-S#

References (*##-##*)