Habitual Control of Goal Selection in Humans

Fiery Cushman and Adam Morris

Department of Psychology, Harvard University

**Abstract**

Goal-directed planning is a hallmark of human behavior. Yet, formal models of goal selection show that it often carries severe computational costs. A key challenge is to understand how humans 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.

**Introduction**

The distinction between habitual and planned action is fundamental to behavioral research (1-4). Habits enable computationally efficient decision making, but at the cost of behavioral flexibility. They form as stimulus-response pairings are “stamped in” following reward, as in Thorndike’s law of effect (3). Planning, in contrast, enables more flexible and productive decision-making. It accomplishes this by searching over a causal model that links candidate actions to their expected outcomes, ultimately selecting actions as a function of their anticipated rewards. However, it carries a severe computational cost: The number of candidate plans to be evaluated rapidly becomes intractable as the complexity of the model grows.

Much research aims to find computationally tractable forms of planning for complex tasks. One important result is that computational costs can be partially mitigated when planning is organized around a task-appropriate goal (5, 6). The resulting goal-specific policies are then available for reuse and recombination, further reducing computation in future planning episodes (7). Recombination is especially productive when goals are organized hierarchically (e.g., turn on the machine to grind the beans to make the coffee) (8). Thus, it is not surprising that psychological research shows that human planning makes widespread use of hierarchical goal structure (3, 9, 10).

In one crucial respect, however, planning around hierarchical goals maintains severe computational constraint: Although planning is easier with a goal, first an appropriate goal must be selected. Suppose, for instance, that your goal is to make a cup of coffee. What is an appropriate sub-goal to select? In principle, an infinite number of possibilities might be entertained, and each one evaluated for its long-term utility. Of course, this merely recapitulates the dilemma of planning, which is that exhaustive search is not feasible. How, then, do humans efficiently alight upon an appropriate sub-goal: getting ground beans?

One potential solution to this problem is to exploit the computational efficiency of habit learning for the purposes of goal selection (11-13). Traditionally, habitual action is modeled as a learned association between a perceptual stimulus and motor response (CITE). This form of learning can be extended, however, to the relation between superordinate goals (an internally represented “stimulus”) and subordinate goal selection (a cognitive “response”), guided by the context-dependent reward history. Thus, for instance, the goal of getting ground beans might be “stamped in” due to the history of reward associated with this goal in past coffee-making episodes. In contrast to many past treatments of the distinction between habit and planning, which treat these mechanisms as competitors for behavioral control (14, 15), this proposal posits a form of co-dependence (see also 16).

In order to provide a detailed model of habitual control over goal selection, and to link it with experimental evidence, we adopt a formal account of habitual versus goal-directed action derived from the reinforcement learning (RL) framework (17). Model-based RL maintains an explicit causal model of the world and uses it to choose actions by assessing their likely consequences. Thus, it enables goal-directed planning. In contrast, model-free RL does not maintain an explicit causal model, and therefore does not allow planning. Rather, it assigns value to candidate actions based on their context-dependent history of reward. The resulting cached policies (akin to stimulus-response habits) are globally adaptive, but may exhibit local irrationality (18, 19).

RL models are widely used in cognitive research because they capture several core features of learning and choice, including in humans (1, 15, 19). Elements of model-free RL, including prediction-error updating and temporal difference learning, are implemented in the midbrain dopamine system (20-22). Human behavior also relies extensively on model-based planning towards goals, which depends on diverse cortical and subcortical regions (4, 19, 23-25).

We use the RL framework to formalize and test for model-free control over goal selection. Once a goal has been selected, subsequent planning to achieve it could 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 captures the notion of a “habit of thought”: Model-free control can contribute to the effective deployment of model-based cognitive routines that facilitate productive and flexible cognition. Consistent with this proposal, recent research emphasizes the pervasive role of model-free control in related elements of higher-level cognition (26, 27), including the gating of working memory (28) and the construction of hierarchical task representations (7). Collectively, such models offer an appealing functional explanation for the neuronal connections between striatum and frontal cortex (29).

The possibility of habitual control over goal selection complements several existing models, both in RL and psychology. Some past computational models of RL directly implement model-free control over hierarchical goal selection, and with promising results (12). This formal approach to model-free control over model-based planning has not, however, been directly linked to experimental research in humans. Meanwhile, psychological models of hierarchical planning recognize the problem of goal selection and have implemented a number of solutions, varying in scope and specificity. These include the use of hidden-layer back propagation networks (9), Pavlovian search heuristics (30), procedural learning mechanisms (31), the chunking of action sequences (32), and other dedicated or domain-specific solutions (3, 33). Here, we pursue two aims: to explicitly link a formal model of goal selection to experimental research in humans, and to use this link in order to provide direct evidence for the role of a specific, well-characterized psychological mechanism: habit learning.

Our task is adapted from a multistep choice paradigm used in prior research (19). The original paradigm behaviorally dissociates the influence of habitual (model-free) and goal-directed (model-based) control on choice. The key feature of this task is that it exploits low-probability connections between behavior and reward. A mechanism employing model-free methods is sensitive to such rewards, stamping in the participant’s prior choice. In contrast, a model-based mechanism planning over a known causal model of the task would discount the link between such actions and rewards because of their known low probability of occurrence. By giving participants repeated choice opportunities, the influence of model-free and model-based control can be dissociated. Several lines of convergent evidence support the alignment of these mechanisms with habitual and goal-directed control, including functional neuroimaging (19), transcranial magnetic stimulation (34), and manipulations of cognitive load (14) and stress (35), among others (36) (CITE MORE).

We modified this task so that it can 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 actions drawn from the set [1,2,3,4]. These choices trigger stochastic transitions 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 Stage 2 state. Participants received detailed instructions and practice trials, including information about the stochastic transitions between Stage 1 and Stage 2.



**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 history (19). 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).

**Results**

*Experiment 1*

We assessed choice on critical trials by comparing instances when the participant experienced reward vs. punishment on the preceding setup trial (i.e. following low-probability transition to the green state). Consistent with our prediction, the mean proportion of trials on which participants selected the shared-goal action following positive reward (89%) was significantly greater than the proportion following negative reward (69%) *t*(134)=-12.5, *p*<.0001 (Fig. 2A).



**Fig. 2.** Bars represent the proportion of trials on which participants chose the shared-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.

In order to capture trial-by-trial variation in the magnitude of the reward obtained on the setup trial, we regressed choice on the model-free goal value[[1]](#footnote-1) using a logistic mixed-effects model, estimating both random intercepts and random slopes at the subject level[[2]](#footnote-2). All mixed-effects analyses were conducted in R (37), making use of the lme4 linear mixed effects package (38). The model-free goal regressor significantly predicted choice (*β* = .191; Wald z-test, z = 12.1, p < .0001). The model was preferred to a null model without the reward (Likelihood ratio test, *χ*2(2) = 266.0, p < .0001). In a parametric bootstrap analysis, 0 out of 1000 randomly resampled null models had a likelihood as large as the full model.

Next, we re-estimated this model while including additional regressors for model-based and model-free action values[[3]](#footnote-3). In this analysis the model-free goal regressor again significantly predicted choice (*β* = .200, z = 12.3, p < .0001). The model was preferred to the null model (*χ*2(4) = 298.2, p < .0001). In a bootstrap analysis, 0 out of 1000 randomly resampled null models had a likelihood as large as the full model. The model based action predictor (*β* =.221, z = 7.3, p < .0001) was significant, and the model-free action predictor (*β* = .054, z = 1.87, p = .062) predictor had a trending effect.

*Experiment 2*

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 shared-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 critical trials: After a low-probability transition to the green triangle state, participants were more likely to choose the shared-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).

Following our analytic approach in Experiment 1, in a mixed effects model on same-type trials the model-free goal regressor was significant (*β* = .056, z = 4.51, p < .0001). The model was preferred to a null model (*χ*2(2) = 27.8, p < .0001; by bootstrapping, p < .001). This effect remained after controlling for model-based and model-free action values, as in Experiment 1.

In the simple mixed-effects model on the different-type goal trials, the model-free goal regressor was not significant (*β* = .009, z = .784, p = .433). The model was not preferred to a null model (*χ*2(2) = .615, p = .74; by bootstrapping, p = .55). We also estimated a model with both same- and different-type critical trials, which included the model-free goal value and an interaction between that value and the trial type. In that model, the interaction was significant (*β* = .049, z = 2.62, p < .01), and the model was preferred to a null model with the interaction term removed (*χ*2(4) = 10.7, p < .05; by bootstrapping, p < .01). Same-type trials were coded as 1 and different-type trials were coded as 0, so the positive interaction term indicates that the model-free goal effect was significantly stronger for same-type trials.



**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.

*Experiment 3*

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 (8, 39). 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 shared-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 shared-goal action following positive reward (85%) than following negative reward (69%) *t*(172)=-9.17, *p*<.0001 (Fig. 2C).



**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.

In a mixed-effects model the model-free goal regressor significantly predicted choice (*β* = .143, z = 9.62, p < .0001), and the model was preferred to a null model (*χ*2(2) = 238.1, p < .0001; by bootstrapping, p < .001). This effect remained after controlling for model-based and model-free action values, as in Experiments 1 and 2.

The results of Experiment 3 speak against an alternative interpretation of Experiment 1 according to which statistical structure of state transitions could support associative pairings between shared-goal Stage 1 options, and thus associative transfer of reward values (15). 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.

*Computational Model*

To validate our analytic approach we specified a computational model of learning and choice, including traditional model-based and model-free control along with model-free goal learning. We used this computational model to generate simulated data for Experiment 2, and showed that our observed results were obtained if and only if the computational model included model-free goal learning.

The game was implemented as a Markov decision process with six states: The initial Stage 1 state, and five Stage 2 states. The Stage 1 state had four possible actions (i.e. the four numbers), only two of which were available on any given trial. The Stage 2 states had only one possible action (i.e. clicking on the object), which led to a reward. The rewards were randomly generated for each agent by the same process as in the behavioral tasks.

The agents had three learning mechanisms. Their model-free reinforcement learning mechanism was the SARSA algorithm with eligibility traces (5). Agents estimated a model-free value of the state-action pair (*s,a*), denoted *MFV(s,a)*. In Stage 1, agents chose an action *a* and transitioned to state *s.* The value update for *MFV(1,a)* occurs by temporal difference learning with learning rate :



In Stage 2, agents chose the only available action *a’* (i.e. clicking on the object) and received reward *r*. Again, value update is given by temporal difference learning:



In keeping with prior computational models of stochastic sequential decision-making paradigms (*6,22*), we also implemented an update of Stage 1 value representations following reward by applying an eligibility trace :



Agents’ model-based learning mechanism implemented a basic forward planning technique. Agents maintained a model-based value of each state-action pair, denoted *MBV(s,a)*. We assumed knowledge of the trial-type-dependent reward distributions on the part of the model-based controller. Thus, we separately indexed Stage 2 states according to the relevant trial type. To calculate the model-based value of each action from state 1, agents estimated the transition probability from *a* to *s*, denoted *T(a,s)*, by counting the number of observed transitions from *a* to *s* and normalizing the counters. Counters were initialized to 1, yielding a symmetric prior distribution over transition probabilities under the Dirichlet model. Then:



After transitioning to state *s*, performing the only available action *a’*, and receiving reward *r*, the model-based update was:



The third learning mechanism was our proposed mechanism, model-free learning on goal selection. After a trial with chosen action *a* and received reward *r*, agents inferred the intended goal *g(a)* by:



Agents then updated the model-free value of the goal, *MFG(g(a))*, by:



To determine the probability of selecting action *a* out of choice set (*a*,*b*), agents took a weighted average *Wa* of the three values and entered it into a softmax function:



Thus, agents were characterized by five parameters: (the learning rate), (the eligibility trace), (the softmax temperature), (the model-based weight), and (the model-free weight).



In each simulation, 200 agents were generated with parameters sampled uniformly over plausible ranges[[4]](#footnote-4). We ran two simulations: one where agents performed model-free goal learning, and one where they did not. We then analyzed the agents’ behavior by the same process as in the behavioral tasks.

In the simulation with model-free goal learning, on same-type trials agents chose the shared-goal action 66.3% of the time after a reward and 51.2% of the time after a punishment (*t*(199) = -.694, p < .0001). The simple mixed-effects model on same-type trials estimated a model-free goal coefficient of .081 (z = 7.35,p < .0001), and was preferred to a null model (*χ*2(2) = 63.1, p < .0001). The complete mixed-effect model showed similar results.

On different-type trials, agents chose what would have been the shared-goal action 48.6% of the time after a reward and 47.6% of the time after a punishment (*t*(186) = -.292, p = .77). The simple mixed-effect model estimated a model-free goal coefficient of .0098 (z = 0.649,p = .516), and was not preferred to a null model (*χ*2(2) = .421, p = .81). The model that combined same- and different-type critical trials showed a significant interaction between model-free goal value and critical trial type (*χ*2(3) = 14.4, p < .005).

In the simulation where agents did not perform model-free goal learning (i.e., ) agents showed no difference in behavior following a reward versus a punishment on same-type trials (*t*(199) = .71, p = .481). The simple mixed-effects model on same-type trials was not preferred to a null model (*χ*2(2) = .483, p = .786). Neither was the complete model (*χ*2(4) = 2.33, p = .675). The model combining same- and different-type trials did not show a significant interaction effect (*χ*2(3) = 0, p = 1).



**Discussion**

We find that goal selection in humans 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. In our simple task this mechanism appears suboptimal, because participants could easily have performed an exhaustive search over candidate goals and thereby attained a higher average rate of return. Yet, the same mechanism may be crucial in allowing humans to avoid the computational burden of full model-based evaluation for the kinds of complex tasks that we routinely face in everyday life—tasks for which exhaustive search over candidate goals is computationally intractable.

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 known empirical phenomena that blend canonical features of habits and goals. Contextual cues can trigger goal pursuit outside of conscious awareness (40), 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 (41). 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 (42). 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 (43, 44).

Although habitual control over goal selection ameliorates some of the computational demands of hierarchical planning, there is no free lunch: By relying on model-free value representation to select goals, an agent forgoes the opportunity for optimal planning. This is apparent in our own task, where the influence of model-free goal values tends to reduce participants’ payoff on critical trials, compared with goal values derived from model-based evaluation. Thus, a key task for any actor is to optimally balance the efficacy of model-free influence against the productivity of model-based influence. Current research offers several promising avenues for understanding how humans accomplish this task (45-48).

Within the context of our own model, one way to fine-tune the balance between efficiency and productivity would be to allow for the selection and evaluation of more than one candidate sub-goal. In the model that we implement, only a single subgoal is retrieved and then it is deterministically adopted. A trivial extension of this model would retrieve multiple goals with a probability proportional to their model-free value. Then, each of the candidate goals could be evaluated by model-based means. This mechanism would preserve the possibility of model-based planning not only at the level subsidiary to the selected goal (as we demonstrate here) but also at the level of the selected goal itself. In this case, the function of model-free value assignment would be to restrict the space of subgoals subject to evaluation.

Another key factor determining the utility of habitual goal selection is the accuracy of the model-free goal values. An agent with highly accurate value representations sacrifices little by turning over goal selection to model-free control, while an agent with inaccurate value representations sacrifices much. The experiments that we report establish model-free value representations by manipulating the reward history associated with goal selection. However, obtaining complete and accurate model-free value representations through direct trial-and-error learning will not be feasible for many complex tasks. Critically, past research shows that model-free value representations can be established by a variety of other pathways in humans. For instance, people can generate fictive reward signals by simulating rewards according to an internally specified model (e.g., daydreaming about the bonus they will get if they put in extra hours at work). Such fictive rewards can establish model-free value representations (14), including in the striatum (49). In essence, this uses model-based processes to cache a reward value in memory prior to decision-making, thus avoiding the need for online planning during behavioral control (12, 46).

There is also some evidence that both observational learning and direct instruction in social contexts can establish model-free value representations (50-52). The possibility of cultural transmission of goal/subgoal structure by observational learning or instruction stands out as a particularly likely explanation for the efficiency and power of goal-directed behavior in humans. Cached model-free value assignment to subgoal selection may operate as an important repository for cultural knowledge of this form. This implies a codependence between two capacities that are remarkably developed in humans: cultural transmission (53) and in their capacity for productive and flexible reasoning (54).

[Some concluding remark here, summing up the basic point]

References: *To be copied from endnote appendix at the bottom of this document*

**Acknowledgments:** We thank Michael Frank, Samuel Gershman, Wouter Kool and Josiah Nunziato for their advice and assistance. This research was supported by grant N00014-14-1-0800 from the Office of Naval Research.

**Methods**

*Participants*

A total of 960 subjects were recruited on Amazon Mechanical Turk to participate in a two-stage Markov decision task. Subjects were excluded from analysis if they timed out on more than 50 trials. Following past research on a similar task (24), we also excluded subjects who did not meet a minimum threshold of learning. We ran a Monte Carlo simulation of 10,000 agents performing the task randomly, and determined the 95th percentile of their final scores. We excluded subjects whose final scores were below this cutoff. After applying our exclusionary criteria, there were 135 subjects and 3806 critical trials in Experiment 1, 303 subjects with 4231 same-type critical trials in Experiment 2 (along with 2137 different-type critical trials), and 173 subjects and 4755 critical trials in Experiment 3.

*Experiment 1 Design*

In Experiment 1, four Stage 1 options (represented by the numbers 1 through 4) each led probabilistically to one of three Stage 2 states (represented by the colors red, blue, and green). 1 and 3 led to red with .8 probability and green with .2 probability, while 2 and 4 led to blue with .8 probability and green with .2 probability. These color states in turn had only one available action, which deterministically led to a reward. Subjects were explicitly told these transition probabilities, and were trained on them in the practice rounds. The high-probability transition of each number became the “goal” of that number – the goal of clicking on 1 would be to get blue, the goal of clicking on 2 would be to get red, etc.

The rewards for each color were initialized uniformly at random on a range of -4 points to +5 points, and varied according to a bounded Gaussian random walk for the remainder of the experiment. After each round, the drift was sampled from a normal distribution with (μ=0, σ=1.8), rounded to the nearest integer, and added to the current reward level[[5]](#footnote-5).

On each trial subjects were presented with only two of the four number options. The option pairs to present were chosen randomly, with the constraint that the high-probability transitions of the two options had to lead to different colors – i.e. 1 and 3 could not be paired. After clicking on one of the two numbers, subjects transitioned to a color, clicked on the color, and received a reward.

Subjects completed 75 practice trials followed by 175 rewarded trials. On the rewarded trials, subjects had only 4 seconds to make their choice between the two numbers. If they did not make a choice within 4s the trial would time out and the next trial would begin.

*Experiment 2 Design*

Experiment 2 was closely modeled on Experiment 1 with a few changes. In Experiment 1, Stage 2 states only varied in their color (blue, red, or green). In Experiment 2, they also varied in their shape. There were three shapes: square, circle, and triangle.

In Experiment 1, reward distribution following Stage 2 was uniquely determined by the color of the Stage 2 option. In Experiment 2, each color and shape had a separate drifting reward distribution, and the reward value of an object could either be determined by its color or shape. On ‘color’ trials, it was the color of the object which determined the reward. On ‘shape’ trials, it was the shape of the object. Subjects were informed of the trial type before each trial. The flow of Experiment 2 is depicted in Figure 3B.

In Experiment 2, our analyses contrasted two types of critical trials: ‘same-type’ trials, which had the same type as the previous trial, and ‘different-type’ trials, which had a different type.

*Experiment 3 Design*

Experiment 3 was also closely modeled on Experiment 1 with a few changes. Before being exposed to the structure of the main task, subjects were trained on a set of intuitive, deterministic transitions from letters to numbers (Figure S5A). After becoming familiar with those transitions, subjects proceeded with the same task as above. All non-critical trials had exactly the same structure as in Experiment 1, with a choice between two numbers leading to a color, which in turn led to one of three drifting reward distributions. However, on critical trials, subjects instead were presented with a choice between two letters. Subjects chose a letter and received a number (in accordance with the transitions in Figure S5A). They then clicked on that number and, in the usual way, got to a color which led them to a reward. The critical trials thus required a goal-directed system to plan one extra step ahead.

1. Dolan RJ & Dayan P (2013) Goals and habits in the brain. *Neuron* 80(2):312-325.

2. Thorndike EL (1898) Animal intelligence: An experimental study of the associative processes in animals. *Psychological Monographs: General and Applied* 2(4):i-109.

3. Norman DA & Shallice T (1986) *Attention to action* (Springer).

4. Balleine BW & Dickinson A (1998) Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* 37(4):407-419.

5. Boyan JA & Moore AW (1996) Learning evaluation functions for large acyclic domains. *ICML*, pp 63-70.

6. Zhang NL & Zhang W (1997) Fast value iteration for goal-directed Markov Decision Processes. *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, (Morgan Kaufmann Publishers Inc.), pp 489-494.

7. Collins AG & Frank MJ (2013) Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review* 120(1):190.

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

9. Botvinick MM (2008) Hierarchical models of behavior and prefrontal function. *Trends in cognitive sciences* 12(5):201-208.

10. Lashley KS (1951) *The problem of serial order in behavior* (Wiley, New York).

11. Mannella F, Gurney K, & Baldassarre G (2013) The nucleus accumbens as a nexus between values and goals in goal-directed behavior: a review and a new hypothesis. *Frontiers in behavioral neuroscience* 7.

12. Sutton RS, Precup D, & Singh S (1999) Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112(1):181-211.

13. Botvinick M & Weinstein A (2014) Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1655):20130480.

14. Otto AR, Gershman SJ, Markman AB, & Daw ND (2013) The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychol Sci* 24(5):751-761.

15. Daw ND & Shohamy D (2008) The cognitive neuroscience of motivation and learning. *Social Cognition* 26(5):593-620.

16. Gershman SJ, Markman AB, & Otto AR (2012) Retrospective Revaluation in Sequential Decision Making: A Tale of Two Systems.

17. Sutton RS & Barto AG (1998) *Introduction to reinforcement learning* (MIT Press).

18. Dickinson A, Balleine B, Watt A, Gonzalez F, & Boakes RA (1995) Motivational control after extended instrumental training. *Learning & behavior* 23(2):197-206.

19. 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.

20. Bayer HM & Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47(1):129-141.

21. McClure SM, Berns GS, & Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. *Neuron* 38(2):339-346.

22. O'Doherty JP, Dayan P, Friston K, Critchley H, & Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. *Neuron* 38(2):329-337.

23. Simon DA & Daw ND (2011) Neural correlates of forward planning in a spatial decision task in humans. *The Journal of neuroscience* 31(14):5526-5539.

24. Gläscher J, Daw N, Dayan P, & O'Doherty JP (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66(4):585-595.

25. Deserno L*, et al.* (2015) Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences*:201417219.

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

27. Graybiel AM (2008) Habits, rituals, and the evaluative brain. *Annu. Rev. Neurosci.* 31:359-387.

28. O'Reilly RC & Frank MJ (2006) Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation* 18(2):283-328.

29. Miller EK (2000) The prefrontal cortex and cognitive control. *Nature Reviews* 1:59-65.

30. Huys QJ*, et al.* (2012) Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology* 8(3):e1002410.

31. Anderson JR (1996) ACT: A simple theory of complex cognition. *American Psychologist* 51(4):355.

32. Dezfouli A, Lingawi NW, & Balleine BW (2014) Habits as action sequences: hierarchical action control and changes in outcome value. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1655):20130482.

33. Cooper R & Shallice T (2000) Contention scheduling and the control of routine activities. *Cognitive Neuropsychology* 17(4):297-338.

34. Smittenaar P, FitzGerald TH, Romei V, Wright ND, & Dolan RJ (2013) Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron* 80(4):914-919.

35. Otto AR, Raio CM, Chiang A, Phelps EA, & Daw ND (2013) Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences* 110(52):20941-20946.

36. Otto AR, Skatova A, Madlon-Kay S, & Daw ND (2014) Cognitive Control Predicts Use of Model-based Reinforcement Learning.

37. Statistical Package R (2009) R: A language and environment for statistical computing. *Vienna, Austria: R Foundation for Statistical Computing*.

38. Bates D, Maechler M, & Bolker B (2012) lme4: Linear mixed-effects models using S4 classes.

39. Ribas-Fernandes JÚJF*, et al.* (2011) A neural signature of hierarchical reinforcement learning. *Neuron* 71(2):370-379.

40. Huang JY & Bargh JA (2014) The Selfish Goal: Autonomously operating motivational structures as the proximate cause of human judgment and behavior. *Behavioral and Brain Sciences* 37(02):121-135.

41. Lhermitte F (1983) ‘Utilization behaviour’and its relation to lesions of the frontal lobes. *Brain* 106(2):237-255.

42. Adamson RE (1952) Functional fixedness as related to problem solving: A repetition of three experiments. *Journal of experimental psychology* 44(4):288.

43. Sfard A (1991) 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):1-36.

44. Perkins DN & Salomon G (1989) Are cognitive skills context-bound? *Educational researcher* 18(1):16-25.

45. 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.

46. Silver D, Sutton RS, & Müller M (2008) Sample-based learning and search with permanent and transient memories. *Proceedings of the 25th international conference on Machine learning*, (ACM), pp 968-975.

47. Daw ND & Dayan P (2014) The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369(1655):20130478.

48. Huys QJ*, et al.* (2015) Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences* 112(10):3098-3103.

49. Lohrenz T, McCabe K, Camerer CF, & Montague PR (2007) Neural signature of fictive learning signals in a sequential investment task. *Proceedings of the National Academy of Sciences* 104(22):9493.

50. Olsson A & Phelps EA (2007) Social learning of fear. *Nature Neuroscience* 10(9):1095-1102.

51. Biele G, Rieskamp J, Krugel LK, & Heekeren HR (2011) The neural basis of following advice. *PLoS Biology* 9(6):e1001089.

52. Doll BB, Jacobs WJ, Sanfey AG, & Frank MJ (2009) Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. *Brain Research* 1299:74-94.

53. Boyd R, Richerson PJ, & Henrich J (2011) The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences of the United States of America* 108(Supplement\_2):10918-10925.

54. Pinker S (2010) The cognitive niche: Coevolution of intelligence, sociality, and language. *Proceedings of the National Academy of Sciences* 107(Supplement 2):8993-8999.

1. We defined the “model-free goal value” as the reward obtained on the setup trial immediately preceding the critical trial. Although formal approaches to model-free reinforcement learning (e.g. Q learning) typically estimate value according to a geometrically-weighted sum of all past rewards9, past experimental research indicates more robust statistical estimates of model-free value assignment under the simplifying assumption that the most recent reward experience dominates value representation. This estimation technique has been used in past studies of stochastic sequential decision-making paradigms4,10. In order to further validate this analytic approach, below we show that it successfully recovers evidence for model-free value assignment to goal selection from the data generated by a formal computational model of our hypothesized mechanism. [↑](#footnote-ref-1)
2. For convergence purposes, all models allowed correlation among random slopes but not between random slopes and the random intercept. [↑](#footnote-ref-2)
3. Consistent with our analytic approach, an action’s “model-free value” was the last reward the agent received from selecting that action, and an action’s “model-based value” was the last reward the agent received from that action’s associated color. These rewards could potentially have been received many rounds ago, and convergence of our model depended upon temporally discounting rewards. We implemented a discounting factor of *γ* = .85. [↑](#footnote-ref-3)
4. was sampled from a uniform distribution from 0 to 1, which we denote as U(0,1). was sampled from U(.5,1). was sampled from U(0,1.5). For the weights, three variables – ,,and – were sampled from U(0,1), and then and .

   

   [↑](#footnote-ref-4)
5. In cases where drift selected a reward level outside the bounds of [-4, 5], the reward would ‘rebound’ by the amount of the excess. [↑](#footnote-ref-5)