# Habitual control of goal selection in humans

Fiery Cushman[1] and Adam Morris

Department of Psychology, Harvard University, Cambridge, MA 20138

Edited by Terrence J. Sejnowski, Salk Institute for Biological Studies, La Jolla, CA, and approved September 10, 2015 (received for review March 31, 2015)

**Humans choose actions based on both habit and planning. Habitual control is computationally frugal but adapts slowly to novel circumstances, whereas planning is computationally expensive but can adapt swiftly. Current research emphasizes the competition between habits and plans for behavioral control, yet many complex tasks instead favor their integration. We consider a hierarchical architecture that exploits the computational efficiency of habitual control to select goals while preserving the flexibility of planning to achieve those goals. We formalize this mechanism in a reinforcement learning setting, illustrate its costs and benefits, and experimentally demonstrate its spontaneous application in a sequential decision-making task.**

planning | goal selection | habit | reinforcement learning | hierarchical control

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 is accomplished by first searching over a causal model linking candidate actions to their expected outcomes and then selecting actions based on their anticipated rewards. Planning imposes a severe computational cost, however, as the size and complexity of a model grows.

Past research emphasizes the competition between habitual and planned control of behavior (5, 6). Habitual control is favored when an individual has extensive experience with a task and when the optimal behavior policy is relatively consistent across time; meanwhile, planning is favored for novel tasks and when the optimal policy is variable, provided that an agent represents an adequate model of their task (7).

Methods of integrating habitual and planned control have received less attention (8–10), yet real-world tasks often favor elements of each. Consider, for instance, a seasoned journalist who reports on new events each day. At a high level of abstraction, her reporting is structured around a repetitive series of goal-directed actions: follow leads, interview sources, evade meddling editors, etc. Because these actions are reliably valuable for any news event, their selection is an excellent candidate for habitual control. The concrete steps necessary to carry out any individual action will be highly variable, however—optimal behavior when interviewing a pop star may be suboptimal when interviewing the Pope. Thus, the implementation of the abstract actions is an excellent candidate for planning. This example illustrates the utility of nesting elements of both habits and plans in a hierarchy of behavioral control (11–13).

Indeed, it is widely recognized that humans mentally organize their behavior around hierarchically organized goals and subgoals (3, 14, 15). In principle, hierarchical organization can be implemented exclusively by habitual control (16), or exclusively by planning (13, 17). However, these homogenous mechanisms foreclose the possibility of tailoring the means of control (habit vs. planning) to the affordances of a particular level of behavioral abstraction. Our aim is to show that humans solve this dilemma by exerting habitual control over the process of goal selection, while using planning to attain the goal once selected.

Traditionally, habits are modeled as a learned association between a perceptual stimulus and motor response. Our proposal entails an extension of habit learning to the relation between superordinate and subordinate goals: a superordinate goal can serve as the internally represented stimulus triggering a cognitive response of subordinate goal selection. Thus, for instance, the goal of getting an interview with a key source might be stamped in due to the history of reward associated with selecting this goal in past news-reporting episodes.

Colloquially, this captures the idea of a "habit of thought": habitual control can contribute to the effective deployment of cognitive routines that facilitate productive and flexible cognition. This proposal is consonant with recent research emphasizing the pervasive role of model-free control in related elements of higher-level cognition (18, 19), including the gating of working memory (20) and the construction of hierarchical task representations (21). These models offer an appealing functional explanation for the neuronal connections between striatum and frontal cortex (22).

## A Reinforcement Learning Perspective

Our proposal can be formalized in the reinforcement learning (RL) setting (23). RL models are widely used in cognitive research because they capture several core features of learning and choice, including in humans (1, 6, 24). We draw especially on two features of RL: the implementation of habitual versus planned action, and the implementation of hierarchical control.

The core principles of habitual and planned control are embodied in two broad classes of RL algorithms. 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 (24, 25). Elements of model-free RL, including prediction-error updating and temporal difference learning, are

## Significance

Human cognition makes widespread use of goal-directed planning. However, exhaustive forward planning for tasks of real-world complexity is prohibitively computationally demanding. Much research aims to find efficient mechanisms for approximate planning. We describe an approach to this problem that exploits the computational efficiency of habit learning to select goal states that are subsequently used in planning. We also provide experimental evidence that humans implement this approach. Our findings illuminate the basis of learning and choice in humans, demonstrate an integration between mechanisms of habitual and planned control, and contribute to our development of computationally tractable planning algorithms.

Author contributions: F.C. and A.M. designed research; F.C. and A.M. performed research; A.M. analyzed data; and F.C. and A.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1] To whom correspondence should be addressed. Email: cushman@fas.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1506367112/-/DCSupplemental.

implemented in the midbrain dopamine system (26–28). Human behavior also relies extensively on model-based planning toward goals, which depends on diverse cortical and subcortical regions (4, 24, 29–31).

Hierarchical control is often accomplished in RL by grouping actions into "options" (12). An option is a sequence of actions (or "policy") bundled collectively for selection by a superordinate controller. For example, tying a bow comprises many individual actions, but these are bundled into a single motor routine. This allows a valuable policy to be generalized across contexts ("policy abstraction"). For instance, learning to tie a bow when putting on one's shoes can generalize to tying a bow while trussing a turkey. In machine learning contexts, intraoption policies are sometimes specified by the programmer; alternatively, they may be learned by model-free methods (16) or by concatenation via repetition into a chunked action sequence (9, 10). These approaches are well suited to situations where the optimal intraoption policy remains constant across episodes, as with tying a bow.

These approaches are poorly suited, however, to circumstances where an intraoption structure is more variable, as when a journalist attempts to secure an interview for a breaking story. Instead, such cases favor intraoption planning toward a goal. Compared with nonhierarchical ("flat") model-based planning, defining options over reliably valuable goals is computationally efficient because it summarizes the expected rewards of implementing the goal, rather than deriving the expected reward from search over a full model of the task (13). For instance, a journalist can retrieve the cached value of pursuing interviews (learned from past experience), rather than deriving the value of these actions by search over many full-length policies for news reporting. Below, we illustrate these computational savings for a specific task.

In summary, modeling habitual goal selection in the reinforcement learning framework comprises three claims. First, hierarchical control can be implemented by defining options over goal states. Second, intraoption policies may be derived from model-based planning toward those goals states. Finally, options may be selected according to cached values derived from model-free update. These features complement tasks where pursuit of a subgoal is reliably valuable (favoring the computational efficacy of model-free valuation of an option), but the means of achieving the subgoal is highly variable (favoring the flexibility of model-based planning within the option).

Our proposal can be contrasted with the recent suggestion that humans may sometimes use goal-directed control in a manner superordinate to habitual action (9, 10)—that is, in which a model-based controller selects habituated action sequences. We seek evidence of the opposite relationship; yet these models are not mutually exclusive. To the contrary, they share the common assumption that humans will flexibly adapt the use of habitual and goal-directed control across levels of hierarchically organized behavior to suit the demands of a particular task. In other words, both propose a "heterarchy" of behavioral control.

The possibility of habitual control over goal selection complements several existing models in RL and psychology (11–13). Some RL algorithms have implemented 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, received a direct experimental test 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 backpropagation networks (14), Pavlovian search heuristics (32), procedural learning mechanisms (33), the chunking of action sequences (9, 10), and other dedicated or domain-specific solutions (3, 34). Here, we aim to explicitly link a formal model of habitual control over goal selection to experimental data.

## Experiment 1

Our task is adapted from a multistep choice paradigm used in prior research (24). The original paradigm behaviorally dissociates the influence of habitual (model-free) and goal-directed (model-based) control on choice. It accomplishes this by exploiting low-probability connections between behavior and reward. A mechanism using model-free value update is sensitive to such rewards, stamping in the participant's prior choice. In contrast, model-based planning over a known causal model of the task discounts the link between actions and reward in such cases according to their low probability of occurrence. By observing participants' choices, 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 (24), transcranial magnetic stimulation (35), and manipulations of cognitive load (5) and stress (36), among others (refs. 37 and 38; but see refs. 9 and 10).

We modified this task to index not only model-free value assignment to actions (as in the original task) but also model-free value assignment to options defined by a goal (Fig. 1*A*). At stage 1 of each trial, participants choose 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 actions that maximize the likelihood of transitioning to the current reward-maximizing stage 2 state. Participants received detailed instructions and practice trials, including both explicit information about the stochastic

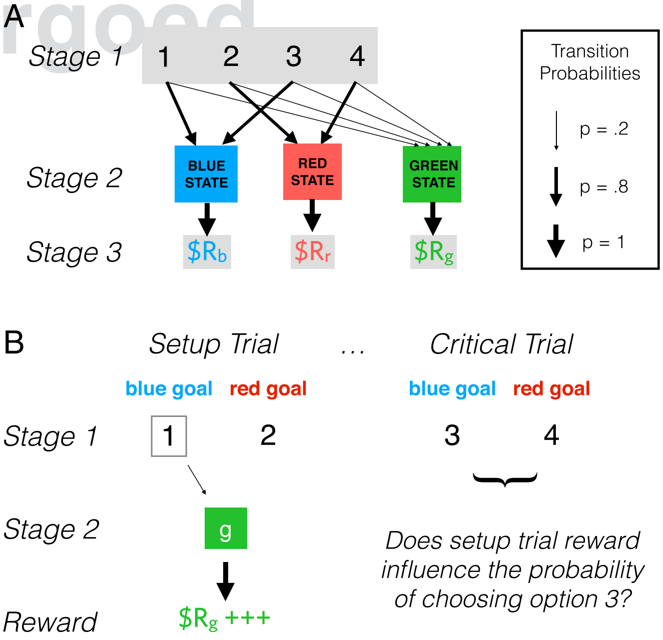

**Fig. 1.** (*A*) In experiment 1a, participants performed a two-stage Markov decision task. They were presented with two possible stage 1 actions drawn from a set of four. These transitioned with variable probabilities to a set of stage 2 actions, which then transitioned deterministically to a set of 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.

transitions between stage 1 and stage 2 and extended practice with those transitions.

Our analysis depends on a subset of trials (Fig. 1*B*). 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 "setup" 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 history of reward following that action (24). 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 "critical" 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 may depend on the assignment of value to their shared goal (getting to blue). In experiment 1, we establish this effect, and in experiment 2 we rule out several alternative explanations of it.

**Experiment 1a.** We assessed choice on critical trials by comparing instances when the participant experienced reward versus 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 (85%) was significantly greater than the proportion following negative reward (69%) [$t_{(216)} = -11.1$, $P < 0.0001$]. Further analysis of these data, and for all subsequent experiments, is presented in *Supporting Information*.

**Experiment 1b.** In experiment 1a, each round of play required a single choice between two actions available at stage 1. This structure does not illustrate the computational savings of model-free value update of options defined over goals. Because each "goal state" (red or blue) deterministically transitions to a single reward distribution, planning toward the goal state is computationally equivalent to planning toward the reward distribution, and is thus no more efficient.

To differentiate subgoals from reward, and to thereby illustrate the computational savings of habitual goal selection, we extend our task to include two sequential rounds of choice (Fig. 2). In this task, when the participant arrives at a colored state in stage 2, they face a choice between two actions. Each action delivers reward from an independent distribution with a drifting mean. Exhaustive model-based search of this decision tree from stage 1 requires the agent to consider transitions and reward distributions for stage 2 actions available from both the blue state and the red state (i.e., actions i–iv). By instead selecting an option defined over a stage 2 goal state, planning over these transitions and rewards is truncated; instead, the expected rewards are now summarized by the value assigned to the option.

Meanwhile, the definition of options over stage 2 goal states enjoys a performance advantage over a flat model-free representation. Specifically, an options-based approach enables the learned value of a stage 2 goal to propagate across stage 1 actions that share high-probability transitions to that goal. In other words, if selecting 1 is rewarding, that value propagates to selecting 3.

In *Supporting Information*, we present formal models implementing pure model-based control, pure model-free control, and our proposed hierarchical integration. We simulate the performance of each model for experiment 1b. We find that pure model-based control attains the highest level of performance (averaging $2.04 in bonus earnings), pure model-free control attains the lowest level (averaging $1.72), and the integrated model attains intermediate performance (averaging $1.97). This occurs because only pure model-based control appropriately discounts the history of reward attained subsequent to the green state, whereas only pure model-free control fails to generalize across stage 1 actions based on their common transition probabilities.

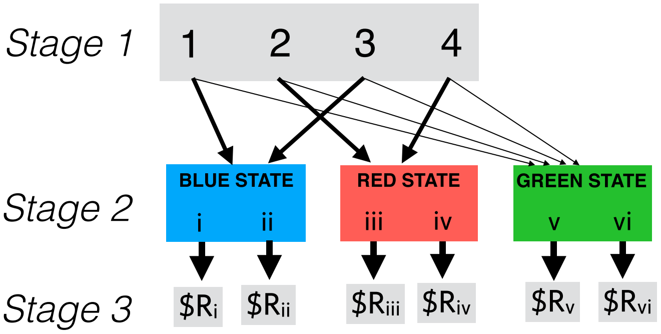

**Fig. 2.** Experiment 1b extends the task introduced in experiment 1a to include a second round of choice. This dissociates putative subgoal states (intermediate colored states) from the terminal states associated with reward.

We then tested a new population of participants on this task. Consistent with our predictions, the mean proportion of trials on which participants selected the shared-goal action following positive reward (79%) was significantly greater than the proportion following negative reward (75%) [$t_{(242)} = -3.2$, $P < 0.005$]. We also fit our formal model to participants' choices in experiment 1b and found that our model was strongly preferred to a null model with the model-free goal selection mechanism removed (exceedance probability = 1). Details are presented in *Supporting Information*.

**Experiment 1: Discussion.** Experiments 1a and 1b provide evidence consistent with habitual control of goal selection. In both experiments, we observed a transfer of learned value across stage 1 actions linked only by their common high-probability transition to a subsequent state. This suggests that participants either engaged in model-based planning over a nonhierarchical (flat) representation the task, or else assigned value to options indexed by the common goal of the subsequent state. However, flat model-based planning is not consistent with the observed influence of reward obtained after low-probability transitions to the green state. Because transitions to this state are equally likely following any stage 1 action, they are irrelevant to planned choice between those actions. Thus, we tentatively conclude that participants adopted model-free hierarchical control over options defined by goal states.

Two concerns limit our confidence in this inference, however. First, it is possible that participants come to represent stage 1 actions with shared high-probability transitions as equivalent for purposes of the task. In other words, they may treat selecting "1" and selecting "3" as the very same action, performing model-free value update on this unified representation. This would account for our results without invoking hierarchical control. A superior experimental paradigm would ensure de novo construction of the intraoption policy using model-based methods on all critical trials. Second, it is possible that participants defined options not over the intermediate goal of attaining a stage 2 option (e.g., "get to red"), but instead over the goal of a terminal state at stage 3 (e.g., "get to Q:11 $R_{iii}$"). This mechanism would still predict a transfer of reward values across stage 1 actions sharing common high-probability transitions, and it would still predict that option values would update based on rewards obtained following low-probability transitions to the green state. Although this alternative shares with our model the premise that options are defined by goal states, and also that intraoption control is accomplished by model-based methods, it is compatible with either model-free or model-based value update of options. This is because the highest level of control reduces to a simple bandit task in which model-based and model-free methods behave equivalently (see also refs. 9 and 10). We designed experiment 2 to target both of these concerns.

## Experiment 2

Experiment 2 involves a sequential decision-making task of similar structure to experiment 1 (Fig. 3). However, in experiment 2, the stage 1 "action" that participants must perform is a mathematical operation. Specifically, in stage 1, participants were presented with a set of three numbers, two of which could be summed to 16, and another two of which could be summed to 21; for example, 7, 9, and 12. By selecting any two numbers that sum to 16, participants deterministically transitioned to one stage 2 state ("state 16"), whereas by selecting any two numbers that sum to 21, they deterministically transitioned to another stage 2 state ("state 21").

We conceive of the abstract action "summation to 16" as an option defined by a goal state. As with the news reporter for whom interviews are always valuable, but must be pursued by variable means, experiment 2 presents participants with a task in which value is restricted to a small number of goals associated reliably with reward (16, 21), but in which goals may be attained by a wide array of actions (i.e., computing the sum of many different pairs of integers).

We limit our analysis to just those trials on which participants are presented with a novel set of numbers at stage 1. This provides a strong safeguard against the possibility of "acquired equivalence" between stage 1 actions. We assume that participants had not acquired an associative equivalence between all possible pairs of integers that sum to 16, or to 21, before the presentation of any given pair in our task.

In addition, experiment 2 alters the structure of the transitions between stage 2 and stage 3 states in a way that allows us to differentiate between options defined over each stage. Crucially, there exists a stage 3 state ($R_{ii}$) that can be deterministically attained via either stage 2 state (16 or 21). An option defined over this terminal reward state ($R_{ii}$) and implementing model-based control would equally favor both summation operations (16 or 21), because both are sufficient to reach the goal. Only when options are defined over the intermediary stage 2 states would reward obtained at $R_{ii}$ systematically bias subsequent selection of a particular stage 1 action.

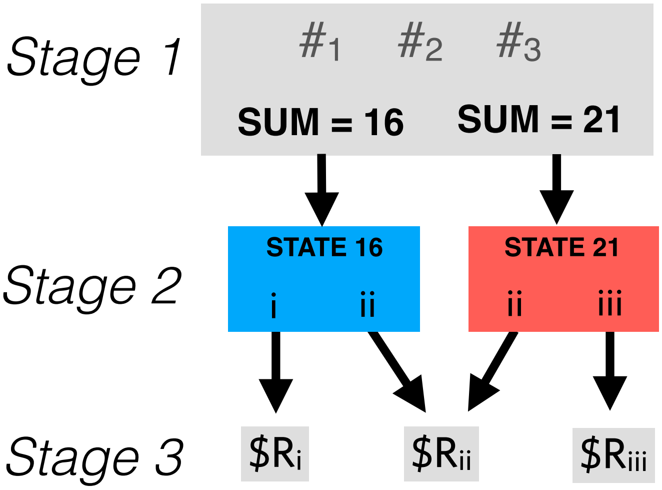

**Fig. 3.** Experiment 2a used a similar structure to experiments 1a and 2b, but with two key differences. First, stage 1 actions were determined by computing the sum of two numbers selected from a total of three presented on each trial. We analyze the subset of critical trials on which participants first encounter a new set of numbers. Second, both stage 2 colored states contained an action that produced a deterministic transition to a single stage 3 reward distribution ($R_{iii}$). This feature dissociates the influence of options defined over stage 2 goals from the influence of options defined over stage 3 goals.

We analyzed data exclusively on critical trials presenting novel sets of numbers at stage 1. Consistent with our prediction, the mean proportion of trials on which participants reselected the same sum goal following positive reward (76%) was significantly greater than the proportion following negative reward (56%) [$t_{(30)} = -2.5$, $P < 0.05$].

**Experiment 2b.** As we have discussed, habitual goal selection af- Q:12 fords computational savings by caching a model-free value representation of goal pursuit—an abstract action defined by a goal. A variant of this proposal uses model-free update to assign value not to the action of pursuing a goal state (i.e., the option "summing to 16"), but instead to the state itself (i.e., state 16). This distinction is subtle but crucial. The latter model could explain the pattern of results obtained in experiment 2a, but without invoking hierarchical control. Rather, control would be implemented by a flat model-based search over a decision tree truncated at stage 2 states, based on values assigned to those states by model-free update.

Critically, this explanation could not apply to experiments 1a and 1b, where value obtained after low-probability transitions to the green state influences subsequent choice. Because transitions to the green state are equally probable from each stage 1 choice, value assigned to the green state cannot influence stage 1 choice. Rather, such an influence implies that value was assigned to the action of pursuing the state as a goal (even if the state is subsequently not attained).

To distinguish between these possibilities in our new task, we adapted the logic from experiment 1 and performed an additional experiment (2b) that implemented low-probability transitions from both summation operations to a third stage 2 state (Fig. 4). This state, indexed "state 0," presents three available actions that deterministically transition to each of the stage 3 states. Our analysis of this task depends on setup trials where participants transition to state 0 and subsequently choose action "ii," which deterministically transitions to a reward distribution available from all stage 2 states ($R_{ii}$). On the subsequent critical trial, participants are presented with a novel set of three numbers at stage 1. We find that they are more likely to repeat the summation goal that they previously selected on the setup trial following reward for state ii (72.3%), compared with punishment [54.6%, $t_{(145)} = -3.8$, $P < 0.001$]. This effect is consistent with model-free valuation of the previously selected option (e.g., summing to 16), but not with valuation of the previously visited stage 2 state (state 0), which is equally available given any action at stage 1.

Why might participants sometimes use a hierarchical task representation that assigns value to options defined by goals (i.e., to goal-directed actions), rather than exclusively relying on a flat task representation that assigns value to truncated branches of a decision tree (i.e., to the corresponding states)? These approaches differ in that the former first selects a goal based on a cached value representation and then searches for a policy to attain it, whereas the latter first searches over potential (truncated) policies, discovering their values by planning. Past research shows that planning algorithms that exploit preselected goals, such as backward reasoning, can attain significant computational savings (39, 40). Savings may be particularly large in real-world domains where the set of possible actions from any given state is very large (e.g., the set of all conceivable actions that a journalist could take when assigned a new article).

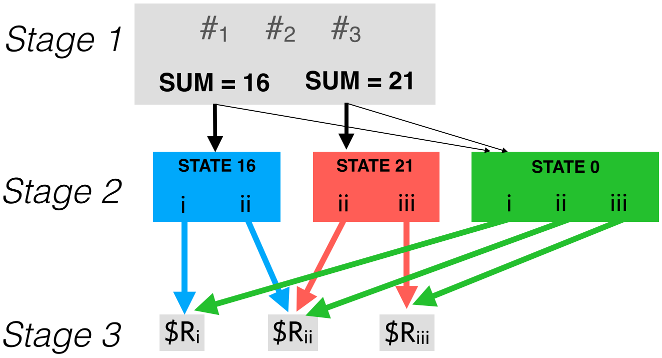

**Fig. 4.** In experiment 2b, a low-probability green state was grafted onto the core decision tree implemented in experiment 2b. This dissociates model-free value update to an option (e.g., "get to state 16") from model-free value update to a state (e.g., state 0) on setup trials involving a low-probability transition from stage 1 to stage 2.

## 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 experiments, this mechanism appears suboptimal, because participants could easily have performed an exhaustive search over candidate goals and thereby attained a higher average rate of reward. However, the same mechanism mitigates the computational burden of full model-based evaluation for the kinds of complex tasks that we routinely face in everyday life.

Although our proposal relies upon the conceptual distinction between habitual (model-free) and planned (model-based) behavioral control, it also demonstrates a mutual dependence between them. This integration captures several empirical phenomena that blend features of habits and goals. Contextual cues can trigger goal pursuit outside of conscious awareness (41), 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 (42). 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 (43). 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 (44, 45).

Habitual goal selection can reduce the computational demands of behavioral control, but there is no free lunch: by relying on habit, an agent forgoes the opportunity for optimal planning. This is apparent in our task, where model-free goal selection reduced participants' payoff, compared with the reward full model-based evaluation could attain. Thus, humans face the challenge of optimally balancing the efficiency of model-free control against the productivity of model-based control. Several promising avenues of research explore how we accomplish this (7, 46–49).

Within the present framework, one approach to fine-tuning this balance is to select and evaluate multiple candidate goals. The model we implemented allows only a single goal to be retrieved and adopted, but a simple extension of this model would retrieve multiple goals with a probability proportional to their model-free value. Then, the value of policies subsequent to each candidate goal state could be evaluated by model-based means (13). In this case, the function of model-free value assignment would be to reduce the size of the planning task, rather than to eliminate it.

The utility of habitual goal selection also depends, of course, on the accuracy of the model-free value representation. An agent with highly accurate representations sacrifices little by turning over goal selection to model-free control, whereas an agent with inaccurate representations sacrifices much. In our experiment, model-free value representations are set by the history of reward. However, obtaining sufficiently accurate representations exclusively by trial and error is not feasible for many complex tasks.

Critically, past research shows that model-free value representations are established by several other means. For instance, value representations can be cached during simulated experience derived from a causal model (8). In addition, both observational learning and direct instruction by social partners establish value representations (50–52). The possibility of cultural transmission of hierarchical goal structure by observational learning or instruction stands out as a likely explanation for the efficiency and power of goal-directed behavior in humans. Cached model-free value assignment to goal selection may serve 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 productive and flexible reasoning (54).

## Methods

**Participants.** A total of 703 subjects was recruited on Amazon Mechanical Turk to participate in Markov decision tasks. Each subject participated in only one task. Subjects gave informed consent, and the study was approved by the Harvard Committee on the Use of Human Subjects.

Subjects were excluded from analysis if they timed out on more than 50 trials, or if their final accumulated score was below zero. After applying our exclusionary criteria, there were 217 subjects and 6,090 critical trials in experiment 1a, 243 subjects and 6,206 critical trials in experiment 1b, 49 subjects and 195 critical trials in experiment 2a, and 194 subjects and 715 critical trials in experiment 2b.

**Experiment Design.** The designs of our experiments are summarized in Figs. 1–4. The two stage 1 options for each trial were always chosen such that the options led to different stage 2 states [i.e., (1, 3) were never paired in experiment 1]. All rewards distributions 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 ($\mu = 0$, $\sigma = 2$), rounded to the nearest integer, and added to the current reward level. In cases where drift selected a reward level outside the bounds of [−4, 5], the reward would "rebound" by the amount of the excess. The rewards on setup trials (those immediately preceding critical trials) were boosted to their extremes by adding +2 or −2 points, depending on the reward distribution's current sign. If the boost selected a reward level outside the bounds, the reward remained at the boundary amount.

After the experiment, participants received a bonus payment based on their accumulated points. Each point was worth 1 cent. Participants were informed of the value of points in the instructions. Each participant completed 75 practice trials followed by 175 rewarded trials. The practice trials were divided into three sections of 25 practice trials each. Sections were designed to ease participants into the task by introducing one task element at a time. On the rewarded trials, subjects had only 4 s to make their choice between the two numbers. If they did not make a choice within 4 s, the trial would time out and the next trial would begin. Practice trials had no time limit. Participants in experiments 1a and 1b saw 26 critical trials each. The spacing of critical trials in experiments 1a and 1b was chosen randomly, with the constraint that they had to be at least three trials apart from each other.

**ACKNOWLEDGMENTS.** We thank Michael Frank, Samuel Gershman, Wouter Kool, Josiah Nunziato, and members of the Moral Psychology Research Laboratory for their advice and assistance. This research was supported by Grant N00014-14-1-0800 from the Office of Naval Research.

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. *Psychol Monogr* 2(4):i–109.
3. Norman DA, Shallice T (1986) Attention to action: Willed and automatic control of behavior. *Consciousness and Self-Regulation*, Advances in Research and Theory, eds Davidson RJ, Schwartz GE, Shapiro D (Springer, New York), Vol 4, pp 1–18. Q:13

4. Balleine BW, Dickinson A (1998) Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology* 37(4-5):407–419.

5. 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.

6. Daw ND, Shohamy D (2008) The cognitive neuroscience of motivation and learning. *Soc Cogn* 26(5):593–620.

7. 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.

8. Gershman SJ, Markman AB, Otto AR (2014) Retrospective revaluation in sequential decision making: A tale of two systems. *J Exp Psychol Gen* 143(1):182–194. Q:14

9. Dezfouli A, Lingawi NW, Balleine BW (2014) Habits as action sequences: Hierarchical action control and changes in outcome value. *Philos Trans R Soc Lond B Biol Sci* 369(1655):20130482.

10. 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.

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. *Front Behav Neurosci* 7:135.

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

13. Botvinick M, Weinstein A (2014) Model-based hierarchical reinforcement learning and human action control. *Philos Trans R Soc Lond B Biol Sci* 369(1655):20130480.

14. Botvinick MM (2008) Hierarchical models of behavior and prefrontal function. *Trends Cogn Sci* 12(5):201–208.

15. Lashley KS (1951) The problem of serial order in behavior. *Cerebral Mechanisms in Behavior*, ed Jeffress LA (Wiley, New York), pp 112–136. Q:15

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

17. Jong NK, Stone P (2008) Hierarchical model-based reinforcement learning: R-max+ MAXQ. *Proceedings of the 25th International Conference on Machine Learning* (ACM, New York), pp 432–439. Q:16

18. Dayan P (2012) How to set the switches on this thing. *Curr Opin Neurobiol* 22(6): 1068–1074.

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

20. O'Reilly RC, Frank MJ (2006) Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput* 18(2):283–328.

21. Collins AG, Frank MJ (2013) Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychol Rev* 120(1):190–229.

22. Miller EK (2000) The prefrontal cortex and cognitive control. *Nat Rev Neurosci* 1(1): 59–65.

23. Sutton RS, Barto AG (1998) *Introduction to Reinforcement Learning* (MIT Press, Cambridge, MA).

24. 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.

25. Dickinson A, Balleine B, Watt A, Gonzalez F, Boakes RA (1995) Motivational control after extended instrumental training. *Learn Behav* 23(2):197–206.

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

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

28. 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.

29. Simon DA, Daw ND (2011) Neural correlates of forward planning in a spatial decision task in humans. *J Neurosci* 31(14):5526–5539.

30. 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.

31. Deserno L, et al. (2015) Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proc Natl Acad Sci USA* 112(5):1595–1600.

32. Huys QJ, et al. (2012) Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Comput Biol* 8(3):e1002410.

33. Anderson JR (1996) ACT: A simple theory of complex cognition. *Am Psychol* 51(4): 355–365. Q:17

34. Cooper R, Shallice T (2000) Contention scheduling and the control of routine activities. *Cogn Neuropsychol* 17(4):297–338.

35. 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.

36. Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND (2013) Working-memory capacity protects model-based learning from stress. *Proc Natl Acad Sci USA* 110(52): 20941–20946.

37. Otto AR, Skatova A, Madlon-Kay S, Daw ND (2015) Cognitive control predicts use of model-based reinforcement learning. *J Cogn Neurosci* 27(2):319–333. Q:18

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

39. Boyan JA, Moore AW (1996) Learning evaluation functions for large acyclic domains. *ICML-96* (Morgan Kaufmann Publishers, San Francisco), pp 63–70. Q:19

40. 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, San Francisco), pp 489–494. Q:20

41. Huang JY, Bargh JA (2014) The selfish goal: Autonomously operating motivational structures as the proximate cause of human judgment and behavior. *Behav Brain Sci* 37(2):121–135.

42. Lhermitte F (1983) "Utilization behaviour" and its relation to lesions of the frontal lobes. *Brain* 106(Pt 2):237–255.

43. Adamson RE (1952) Functional fixedness as related to problem solving; a repetition of three experiments. *J Exp Psychol* 44(4):288–291.

44. Sfard A (1991) On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educ Stud Math* 22(1):1–36.

45. Perkins DN, Salomon G (1989) Are cognitive skills context-bound? *Educ Res* 18(1): 16–25.

46. Pezzulo G, Rigoli F, Chersi F (2013) The mixed instrumental controller: Using value of information to combine habitual choice and mental simulation. *Front Psychol* 4:92.

47. 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, New York), pp 968–975. Q:21

48. Daw ND, Dayan P (2014) The algorithmic anatomy of model-based evaluation. *Philos Trans R Soc Lond B Biol Sci* 369(1655):20130478.

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

50. Olsson A, Phelps EA (2007) Social learning of fear. *Nat Neurosci* 10(9):1095–1102.

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

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

53. Boyd R, Richerson PJ, Henrich J (2011) The cultural niche: Why social learning is essential for human adaptation. *Proc Natl Acad Sci USA* 108(Suppl 2):10918–10925.

54. Pinker S (2010) Colloquium paper: the cognitive niche: Coevolution of intelligence, sociality, and language. *Proc Natl Acad Sci USA* 107(Suppl 2):8993–8999.

55. R Development Core Team (2009) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna). Q:22

56. Bates D, Maechler M, Bolker B (2012) lme4: Linear mixed-effects models using S4 classes. Q:23

57. Burnham KP, Anderson DR (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer Science and Business Media, New York). Q:24

58. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. *Neuroimage* 46(4):1004–1017.

59. McFadden D (1974) Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, ed Zarembka P (Academic, New York), pp 105–142. Q:25

# AUTHOR QUERIES

## AUTHOR PLEASE ANSWER ALL QUERIES

Q: 1_Please contact PNAS_Specialist.djs@sheridan.com if you have questions about the editorial changes, this list of queries, or the figures in your article. Please include your manuscript number in the subject line of all email correspondence; your manuscript number is 201506367.

Q: 2_Please (i) review the author affiliation and footnote symbols carefully, (ii) check the order of the author names, and (iii) check the spelling of all author names, initials, and affiliations. Please check with your coauthors about how they want their names and affiliations to appear. To confirm that the author and affiliation lines are correct, add the comment "OK" next to the author line. This is your final opportunity to correct any errors prior to publication. Misspelled names or missing initials will affect an author's searchability. Once a manuscript publishes online, any corrections (if approved) will require publishing an erratum; there is a processing fee for approved erratum.

Q: 3_Please review and confirm your approval of the short title: Habitual control of goal selection. If you wish to make further changes, please adhere to the 50-character limit. (NOTE: The short title is used only for the mobile app and the RSS feed.)

Q: 4_Please review the information in the author contribution footnote carefully. Please make sure that the information is correct and that the correct author initials are listed. Note that the order of author initials matches the order of the author line per journal style. You may add contributions to the list in the footnote; however, funding should not be an author's only contribution to the work.

Q: 5_Your article will appear in the following section of the journal: Social Sciences (Psychological and Cognitive Sciences). Please confirm that this is correct.

Q: 6_You have chosen the open access option for your paper and have agreed to pay an additional $1350 (or $1000 if your institution has a site license). Please confirm this is correct and note your approval in the margin.

Q: 7_Please verify that all supporting information (SI) citations are correct. Note, however, that the hyperlinks for SI citations will not work until the article is published online. In addition, SI that is not composed in the main SI PDF (appendices, datasets, movies, and "Other Supporting Information Files") have not been changed from your originally submitted file and so are not included in this set of proofs. The proofs for any composed portion of your SI are included in this proof as subsequent pages following the last page of the main text. If you did not receive the proofs for your SI, please contact **PNAS_Specialist.djs@sheridan.com**.

Q: 8_Please check the order of your keywords and approve or reorder them as necessary. Note that PNAS allows up to five keywords; please do not add new keywords unless you wish to replace others.

Q: 9_Quotation marks are used at the first appearance of a word or phrase; they do not appear with each subsequent use of that word or phrase.

# AUTHOR QUERIES

## AUTHOR PLEASE ANSWER ALL QUERIES 2

Q: 10_Per journal style, italics may not be used for emphasis. Adjusted here (“superordinate” in sentence beginning “Our proposal can be contrasted. . ..”) and elsewhere, as applicable, in the text.

Q: 11_Please check all math, equations, and variables in the text for accuracy in typesetting, including typeface (roman or italic), letters, numbers, and symbols to ensure they are complete and displaying properly.

Q: 12_*Experiment 2b* is the only level 2 heading under *Experiment 2*. Please add a second level 2 heading or delete the single one in this section.

Q: 13_For ref. 3 (Norman and Shallice): This reference has been updated. Please confirm title of chapter, title of book, title of series, list of editors, location of publisher, volume number, and page range.

Q: 14_For ref. 8 (Gershman et al.): This reference has been updated (per PubMed). Please confirm year of publication, title of journal, volume number, issue number, and page range.

Q: 15_For ref. 15 (Lashley): This reference has been updated. Please confirm title of book, name of editor, and page range of chapter.

Q: 16_For ref. 17 (Jong and Stone): This reference has been updated. Please confirm location of publisher.

Q: 17_For ref. 33 (Anderson): This reference has been updated. Please confirm that page range is correct.

Q: 18_For ref. 37 (Otto et al.): This reference has been updated (per PubMed). Please confirm year of publication, title of journal, volume number, issue number, and page range.

Q: 19_For ref. 39 (Boyan and Moore): This reference has been updated. Please confirm title of book, name of publisher, and location of publisher.

Q: 20_For ref. 40 (Zhang and Zhang): This reference has been updated. Please confirm location of publisher.

Q: 21_For ref. 47 (Silver et al.): This reference has been updated. Please confirm location of publisher.

Q: 22_For ref. 55 (R Development Core Team): This reference has been updated. Please confirm author.

Q: 23_For ref. 56 (Bates et al.): This information in this reference is incomplete. Please provide additional information (as applicable): name of publisher and location of publisher (city and state/country), or URL address and date on which URL was accessed (month, day, and year).

Q: 24_For ref. 57 (Burnham and Anderson): This reference has been updated. Please confirm location of publisher.

Q: 25_For ref. 59 (McFadden): This reference has been updated. Please confirm location of publisher.

# Supporting Information

## Cushman and Morris 10.1073/pnas.1506367112

### Analysis of Behavioral Data by Logistic Regression

To capture the effect of trial-by-trial variation in the setup trial reward magnitude on choice in the critical trial, we regressed participants' critical trial choices on the setup trial reward using a logistic mixed-effects model, estimating both random intercepts and random slopes at the subject level. [Following past research (22), this model approximates the value representation of a prediction error update mechanism as the most recently observed reward. In simulations presented below, we validate this approximation.]

All models had a single regressor: the value of the reward obtained on setup trials. The reward regressor was grand mean centered. The dependent variable was participant choice on the subsequent critical trials, coded as 1 if participants selected the shared-goal action, and 0 otherwise. Thus, a positive coefficient indicates that participants were more likely to select the shared-goal action following higher reward on the setup trial.

To achieve convergence, models did not allow correlation between the random slope and random intercept. We determined whether the regressor increased the model's likelihood enough to justify inclusion by calculating a null model with the regressor removed, and comparing models using a likelihood ratio test. All mixed-effects analyses were conducted in R (54), making use of the lme4 linear mixed-effects package (55, 56).

In each experiment, the reward obtained on the setup trial significantly predicted choice. The parameter estimates and significance tests for the mixed-effects models are presented in Table S1. $\beta$ is the coefficient of the reward regressor, $\chi^2$ is the statistic value in the likelihood ratio test, and $P$ is the significance level of the likelihood ratio test.

### Computational Model

Below, we present a computational model of learning and choice that includes model-free goal selection alongside traditional model-based and model-free control. Using this model to generate simulated data in our task, we show that our observed results are obtained only when the model includes model-free goal learning. By comparing our mechanism's performance and computational efficiency to that of traditional mechanisms, we also show that our mechanism balances elements of model-based accuracy with model-free efficiency.

We generated simulated data for experiment 1b. The task is a Markov decision process with 10 states: the initial stage 1 state $S_1$, three stage 2 states $S_{2\text{–}4}$, and six reward states $S_{5\text{–}10}$ (Fig. 1 in main text). At $S_1$, four possible actions exist, but only two of these were available on any given trial. $S_{2\text{–}4}$ each had two available actions, and these led to terminal states, each associated with an independent drifting reward. The rewards were randomly generated for each agent by the same process as in the behavioral tasks.

In our simulation, as in the original experiment, each agent completed 175 trials. Although agents made choices in both stage 1 and 2, we focus our exposition on the stage 1 choice because it uniquely juxtaposes the predictions of the three models we consider.

**Mechanisms.** We implemented model-free goal learning with the options framework (10), a common framework for hierarchical policy abstraction. In our instantiation, an "option" is a flexible policy defined by model-based planning toward a goal state, and which is available for selection by a higher-order controller. We defined two options available in stage 1: one with the goal state of blue (denoted $O_1$), and the other with red ($O_2$). (Choices in stage 2 were made similarly, with an option representing each of the two basic actions available in stage 2 states.)

An agent using these options faces two challenges. It must choose an option, and then select a policy designed to attain the chosen option's goal state. Our proposed mechanism addressed the first challenge by maintaining a model-free value for each option in each state $s$, denoted $V(s{,}O_i)$. The values were initialized to zero. After choosing option $O_i$ and transitioning to state $s'$,

$$V(s,O_i) \leftarrow V(s,O_i) + \alpha\left(r + \max_j V(s',O_j) - V(s,O_i)\right),$$

where $r$ is the received reward and $\alpha$ is a learning rate. We included eligibility traces, so the prediction error was applied to every previously chosen state–option pair in a given trial with decay parameter $\lambda$.

Agents used model-free update to summarize the value of options and select between them, but model-based planning to achieve the goal state defined by a given option. Agents maintained the transition probabilities from each stage 1 action $a$ to each stage 2 state $S_j$, denoted $T(S_1{,}a{,}S_j)$. Because participants were told these probabilities explicitly and had extensive practice with them, we assume that agents know the correct transition probabilities. (Our qualitative results are identical if we instead require agents learn and update the probabilities based on experience.) The controller followed a deterministic, greedy intraoption policy and assigned full probability to the action most likely to transition to the goal state. Formally, agents assigned probability to action $a$ under option $O_i$ according to the following:

$$\text{Prob}(a|O_i) = \begin{cases} 1 \ \ \text{if } a = \underset{x\in A(S_1)}{\text{argmax}}\, T(S_1,x,g_i) \\ \qquad 0, \text{otherwise} \end{cases},$$

where $A(S_1)$ is the set of actions available from $S_1$ on this trial, and $g_i \in \{S_{2\text{–}4}\}$ is the goal state associated with option $O_i$. (Action probabilities under stage 2 options were calculated similarly, using transitions from stage 2 actions to terminating reward states.)

Finally, the model-free goal mechanism combined its option values with its intraoption policy to obtain a value for each action $a$ in each state $s$:

$$Q_{MFG}(s,a) = \sum_{i=1}^{2} V(s,O_i) * \text{Prob}(a|O_i).$$

Conversion from state–option values to state–action values allowed us to model participants' behavior using softmax choice over a mixture of action values specified by each of the three models we consider.

For comparison, agents also implemented a flat model-based controller. (We include a flat, not hierarchical, model-based controller for simplicity of exposition. In our task, a hierarchical model-based controller produces qualitatively identical results to the flat model-based controller that we consider here.) To calculate the value of action $a_1$ in state $S_1$, the model-based controller maintained the stage 1 transition probabilities $T(S_1{,}a_1{,}S_j)$, the set of actions available from each stage 2 state, denoted $A(S_j)$, the transition probabilities from each stage 2 action $a_2 \in A(S_j)$ to each reward state $S_k$, denoted $T(S_j{,}a_2{,}S_k)$, and the current value of each reward state, denoted $V(S_k)$. (The values of reward states were learned by prediction error update after every trial with learning rate $\alpha$.) It then calculated the value of option $a_1$ in state $S_1$ according to the following:

$$Q_{MB}(S_1,a_1) = \sum_{j=2}^{4}\left(T(S_1,a_1,S_j) * \max_{a_2 \in A(S_j)}\left(\sum_{k=5}^{10} T(S_j,a_2,S_k)V(S_k)\right)\right).$$

(Stage 2 action values were assigned by evaluating the inner sum of this equation.) This model-based option evaluation mechanism is more accurate than our model-free mechanism because it partials out any rewards obtained from transitions to the green states. However, it comes at the computational cost of evaluating each possible stage 2 pathway. That cost is minor in our simplified task, but in real-world scenarios it could be prohibitive.

Finally, to ensure that our results could only be the product of model-free learning over options, rather than individual actions, we implemented a traditional model-free action learner. (As above, we include a flat, not hierarchical, model-free action controller. In the task we chose to model, a hierarchical model-free action mechanism could produce qualitatively similar results to our proposed mechanism through learned associations between shared-goal actions in the option-specific policies. Experiments 2a and 2b, which use novel action sets on critical trials, demonstrate a control mechanism that cannot rely on such associations. Therefore, we exclude this possibility from our present model of experiment 1b for simplicity of exposition.) We used Q-learning, a common model of human learning and decision making (21). Agents maintained a value for each state–action pair, denoted $Q_{MF}(s,a)$. After choosing action $a$ in state $s$ and transitioning to state $s'$, agents updated their state–action pair values by temporal difference learning with learning rate $\alpha$:

$$Q_{MF}(s,a) \leftarrow Q_{MF}(s,a) + \alpha\left(r + \max_{a'} Q_{MF}(s',a') - Q_{MF}(s,a)\right).$$

As above, we included eligibility traces with the same decay parameter $\lambda$.

Because agents maintained three separate controllers with different state–action values, we produced a weighted mixture of the state–action values, $Q_W(s,a)$, by the following:

$$Q_W(s,a) = w_{MFG}Q_{MFG}(s,a) + w_{MB}Q_{MB}(s,a) + (1 - w_{MFG} - w_{MB})Q_{MF}(s,a),$$

where $w_{MFG}$ and $w_{MB}$ are the relative weights given to the model-free goal and model-based mechanisms. Agents made final action selections for state $s$ by entering the $Q_W$ values into a softmax function:

$$\text{Prob}(a) = \frac{e^{\beta Q_W(s,a)}}{e^{\beta Q_W(s,a_1)} + e^{\beta Q_W(s,a_2)}},$$

where $\beta$ is a temperature parameter and $a_{1,2}$ are the two available actions in state $s$.

Thus, agents were characterized by five parameters: $\alpha$ (the learning rate), $\lambda$ (the eligibility trace), $\beta$ (the softmax temperature), $w_{MFG}$ (the model-free goal weight), and $w_{MB}$ (the model-based weight). Each agent's parameters were randomly sampled as follows. $\alpha$ was sampled from a uniform distribution from 0 to 1, which we denote as $U(0,1)$. $\lambda$ was sampled from $U(0.5,1)$. $\beta$ was sampled from $U(0.5,1.5)$. For the weights, three variables—$U_1$, $U_2$, and $U_3$—were sampled from $U(0,1)$, and then $w_{MFG} = U_1/\sum_{k=1}^{3} U_k$ and $w_{MB} = U_2/\sum_{k=1}^{3} U_k$. We generated 500 agents per simulation and analyzed agents' behavior by the same process as in the behavioral tasks.

**Results.** In the simulation with model-free goal learning, agents chose the shared-goal action 80.2% (±0.7%) of the time after a reward and 66.7% (±0.7%) of the time after a punishment. The mixed-effects model on same-type trials estimated a model-free goal coefficient of 0.1, and was preferred to a null model [$\chi^2$ (2) = 343.1, $P < 0.0001$]. In contrast, when agents did not perform model-free goal learning ($w_{MFG} = 0$), agents showed no difference in behavior following a reward versus a punishment (71.3% vs. 71.5%). Analysis by mixed-effect models similarly showed null results [$\chi^2$ (2) = 0.422, $P = 0.81$].

We also compared the performances of agents who exhibited only model-free goal selection ($w_{MFG} = 1$), model-based control ($w_{MB} = 1$), or model-free control ($w_{MFG} = w_{MB} = 0$). As predicted, our mechanism accumulated more total reward on the task than a pure model-free mechanism but less than a pure model-based mechanism, suggesting that our mechanism balances the accuracy of model-based approaches with the computational efficiency of model-free approaches (Fig. S1).

## Model Fitting

Our analysis of the effect of setup trial reward on critical trial choice, presented in the main text, suggests that participants spontaneously use model-free control over goal selection. As an additional test of this hypothesis, we fit the above computational model to observed data in experiment 1b. Using MATLAB's *patternsearch* function, we fit the five free parameters individually to every participant by maximum likelihood, each time taking the best out of 25 starts distributed across the parameter space. Parameter estimates and pseudo-$R^2$ values are presented in Table S2. The model fit the data significantly better than a chance model for every participant (likelihood ratio tests, all values of $P < 0.0001$), and our parameter of interest, the model-free goal weight $w_{MFG}$, was distributed significantly above zero (sign test, $Z = 16.0$, $P < 0.0001$).

To determine whether the model-free goal learning mechanism explained enough participant data to justify inclusion in our model, we performed Bayesian model comparison, comparing our model to a null model with $w_{MFG}$ set to zero. By allowing a weighted mixture of the two considered alternative models, this null model accommodates participants who are purely model-based, purely model-free, or any mixture thereof.

We computed the Akaike information criterion (AIC) for each participant as an approximation to the Bayesian model evidence for each model (57), and, following Stephan et al. (58), submitted the individual participant AICs to the spm_BMS routine in SPM8 to calculate the exceedance probabilities of the two models. The results are presented in Table S3. The full model has an exceedance probability of 1, indicating that the model with model-free goal selection definitively provides a better fit to the observed data. These results strongly support the conclusion that participants are using model-free control of goal selection.

We validate this approach by fitting the two models to simulated data. Using the same methods as above, we simulated 100 agents with model-free goal learning and 100 agents with $w_{MFG}$ set to zero. When fitting the former data, we were able to recover the true parameters with sufficient accuracy (correlation between true parameter values and parameter estimates in full model was $r = 0.89$), and Bayesian model comparison indicated that the full model was heavily preferred to the null model (exceedance probability = 1). In contrast, when fitting to the data produced with no model-free goal selection, Bayesian model comparison indicated that the null model was heavily preferred (exceedance probability = 1). These results demonstrate that our model comparison approach would only indicate a preference for the full model in the presence of model-free goal selection, validating the above results.

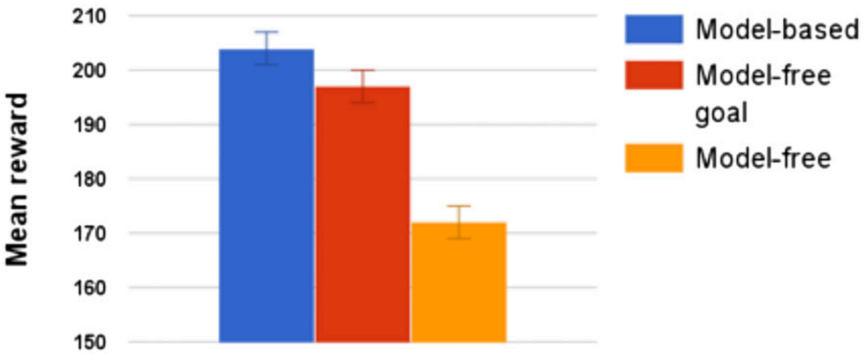

**Fig. S1.** Reward accumulated across 175 trials in experiment 1b by three mechanisms of learning and choice. A pure model-based mechanism, in blue, earned a mean reward of 204 ± 3. A pure model-free mechanism, in yellow, earned 172 ± 3. A model-free goal mechanism, in orange, performed at an intermediate level, earning 197 ± 3.

**Table S1. Parameter estimates and significance tests for the mixed-effects models**

| Experiment | $\beta$ | $\chi^2$ | $P$ |
|---|---|---|---|
| 1a | 0.15 | 309.93 | <0.0001 |
| 1b | 0.032 | 17.7 | <0.001 |
| 2a | 0.21 | 9.85 | <0.01 |
| 2b | 0.1 | 22.3 | <0.0001 |

**Table S2. Parameter estimates for participants in experiment 1b**

| Percentile | $\alpha$ | $\lambda$ | $\beta$ | $w_{MFG}$ | $w_{MB}$ | −LL | Pseudo-$R^2$ |
|---|---|---|---|---|---|---|---|
| 25th | 0.49 | 0.57 | 0.36 | 0.56 | 0.00 | 210 | 0.39 |
| Median | 0.65 | 0.97 | 0.49 | 0.74 | 0.00 | 185 | 0.46 |
| 75th | 0.85 | 1.00 | 0.62 | 0.96 | 0.11 | 158 | 0.54 |

Shown are the 25th, 50th, and 75th percentiles of the distribution of each parameter across subjects. $\alpha$ is the learning rate, $\lambda$ is the eligibility trace decay, $\beta$ is the softmax temperature, $w_{MFG}$ is the relative weight of the model-free goal controller, and $w_{MB}$ is the relative weight of the model-based controller. Also shown are the distribution across subjects of negative log likelihoods and McFadden pseudo-$R^2$ values (59), an approximate measure of the proportion of variance explained by the model.

**Table S3. Model comparison between full model with model-free goal selection, and null model with $w_{MFG}$ set to zero**

| Model | Aggregate −LL | Aggregate AIC | Exceedance probability | Number favoring full model |
|---|---|---|---|---|
| Full | 44223 | 90876 | 1.000 | — |
| Without $w_{MFG}$ | 47194 | 96331 | 0.000 | 198 |

Shown are the aggregate negative log likelihood, aggregate AIC, and exceedance probability for each model. We also report one classical model comparison metric, the number of subjects favoring the full model over the null model by individual likelihood ratio tests (at $P < 0.05$). A total of 198 of 243 subjects favored the full model.

# AUTHOR QUERIES

## AUTHOR PLEASE ANSWER ALL QUERIES

Q: 1_Ref. 56 was uncited and has temporarily been cited with ref. 55 for continuity. Please approve or insert the citation for ref. 56 in its proper sequential location.

Q: 2_Please check all math, equations, and variables in the SI text for accuracy in typesetting, including typeface (roman or italic), letters, numbers, and symbols to ensure they are complete and displaying properly.

Q: 3_Per PNAS style, the use of in-text footnotes is discouraged. Consequently, the footnotes have been removed, and footnote statements have been inserted in parentheses in the SI text. Please check throughout and confirm placement of footnote statements in the SI text.