**Habitual control of goal selection**

Fiery Cushman1 and Adam Morris2

1*Department of Psychology, Harvard University*

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

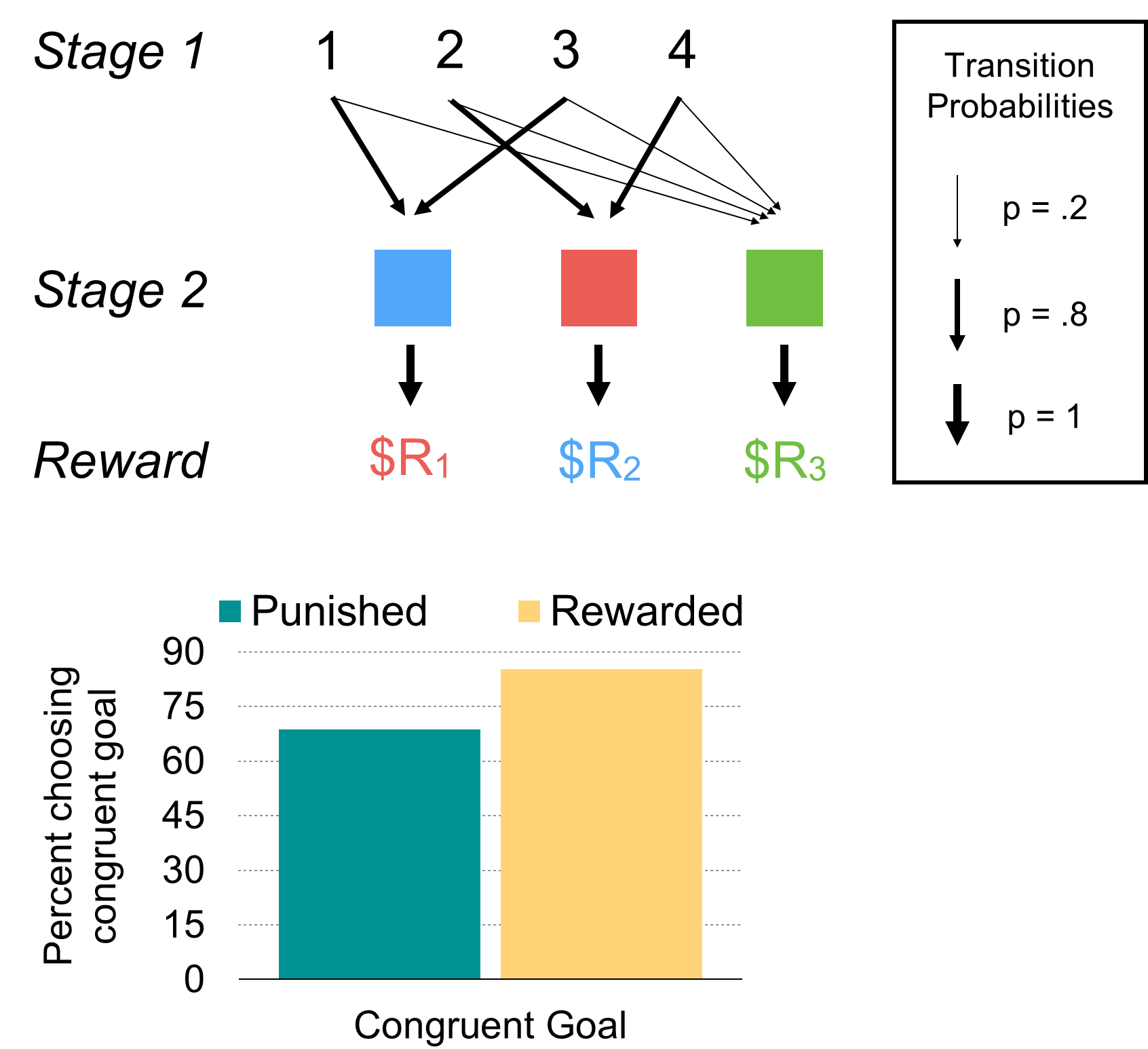
**The distinction between habitual and goal-directed action is fundamental to behavioral research1. Habits form as behaviors are “stamped in” following reward. The resulting stimulus-response pairings enable computationally efficient decision making, but at the cost of behavioral flexibility. In contrast, goal-directed behavior requires planning over a causal model. This enables more flexible decision-making, but at a potentially severe computational cost. Exhaustive search over candidate plans becomes prohibitive as the space of possible goals grows. Thus, a key requirement for goal-directed action is to efficiently select candidate goals with a high likelihood of reward. Here, we provide evidence for a potential solution: Habitual control over the process of goal selection. Although many existing treatments of the distinction between habitual and goal-directed action emphasize their competition over behavioral control2, our results illustrate a codependence between the systems. The role of habitual control in goal-directed action may explain diverse phenomena such as the automatic selection of goals under contextual cuing3, the phenomenon of intrusive goals [C], and the basis of practice effects in controlled cognition [C].**

Our approach depends on a formalization of habitual and goal-directed behavior derived from the reinforcement learning (RL) framework4. Model-based RL maintains an explicit causal model of the world and uses it to choose future actions by assessing their likely consequences. In contrast, model-free algorithms do not maintain an explicit causal model, and thus do not support forward planning. Rather, they assign value to candidate actions based on their context-dependent history of reward. Like Thorndike’s law of effect5, model-free RL increases the probability of rewarded actions and decrease the probability of punished actions. The resulting stimulus-response habits are often globally adaptive, but may exhibit locally irrationality6,7.

RL models capture several core features of learning and choice in humans1,6,8. In particular, the midbrain dopamine reward system appears to implement several key features of model-free RL including prediction-error updating and temporal difference learning9-11. While early research on model-free control emphasized the role of subcortical circuits in selecting motor actions in response to sensory stimuli, recent research emphasizes their pervasive role in higher-level cognition12, including the gating of both short-term13,14 and the long-term15 memory representations. This provides a natural functional account of the widespread neuroanatomical connectivity between striatum and frontal cortex16.

Inspired by formal models of hierarchical RL17, we posit a similar role for model-free control in implementing goal selection. Pure model-based control over goal selection requires an organism to derive the expected value of pursuing candidate goals from a causal model of the rewards obtained during goal pursuit. In order to avoid the computational cost of exhaustive search, an organism could select candidate goals according to the reward history associated with past instances of their selection; i.e., based on model-free value update. Subsequent planning to achieve the selected goal would then proceed in a model-based fashion. In essence, our proposal captures the commonsense notion of a “habit of thought”: the constituent cognitive operations that comprise controlled cognition may themselves be habitized12.

In order to test for the influence of model-free value representations in goal selection we adapted a multistep choice paradigm from prior research6. The original paradigm behaviorally dissociates the influence of habitual (model-free) and goal-directed (model-based) control on choice, and is well-validated2,18-20. Our modification allows us to index not only model-free value assignment to actions (as in the original task) but also model-free value assignment to goals that may be subsequently pursued via model-based planning (Figure 1A). At Stage 1 of each trial participants make a choice between two options drawn from the set (1, 2, 3, 4). These choices stochastically transition to a second set of three states (blue, red, green). Finally, Stage 2 states deterministically transition to three unique reward distributions. The rewards drift over time. Thus, participants are motivated to choose Stage 1 options that maximize the likelihood of transitioning to the current reward-maximizing final state. For task details, see Supplementary Materials.

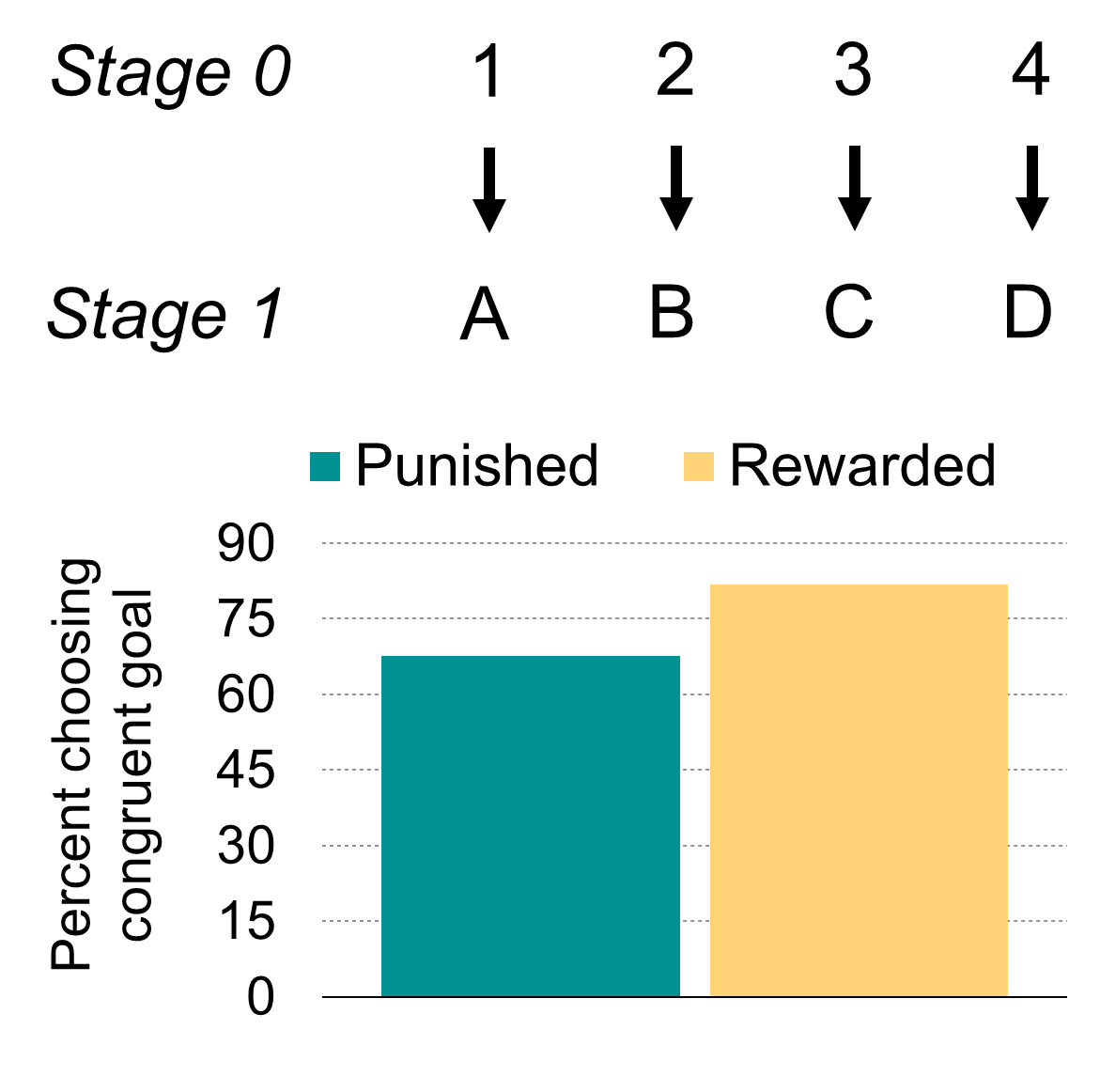


Our analysis depends on a critical subset of trials. For example, a participant is presented with the choice set (1, 2) at Stage 1 and chooses 1. Because 1 typically leads to the blue state, we assume that this participant’s goal was to transition to blue. On our critical trails, however, they experience 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 assignment of value to an action would increase the likelihood of selecting option A on subsequent trials due to the positive reward history6. Our interest, however, is in the model-free assignment of value to a goal; in this case, the goal of transitioning to red. If the experience of reward increases the likelihood of selecting red 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 an value to the specific action (choosing 1); rather, it depends on the assignment of value to their shared goal (getting to red).

We assessed trials of this type by comparing instances when the participant experienced reward vs. punishment following low-probability transition to the green state. Participants were significantly more likely to choose the congruent-goal option following positive reward (85%,SEM=1.2%) than following negative reward (69%,SEM=1.3%) *t*(217)=-11.2, *p*<.0001 (Figure 1B). Supporting analyses for all experiments are presented in supplementary materials.

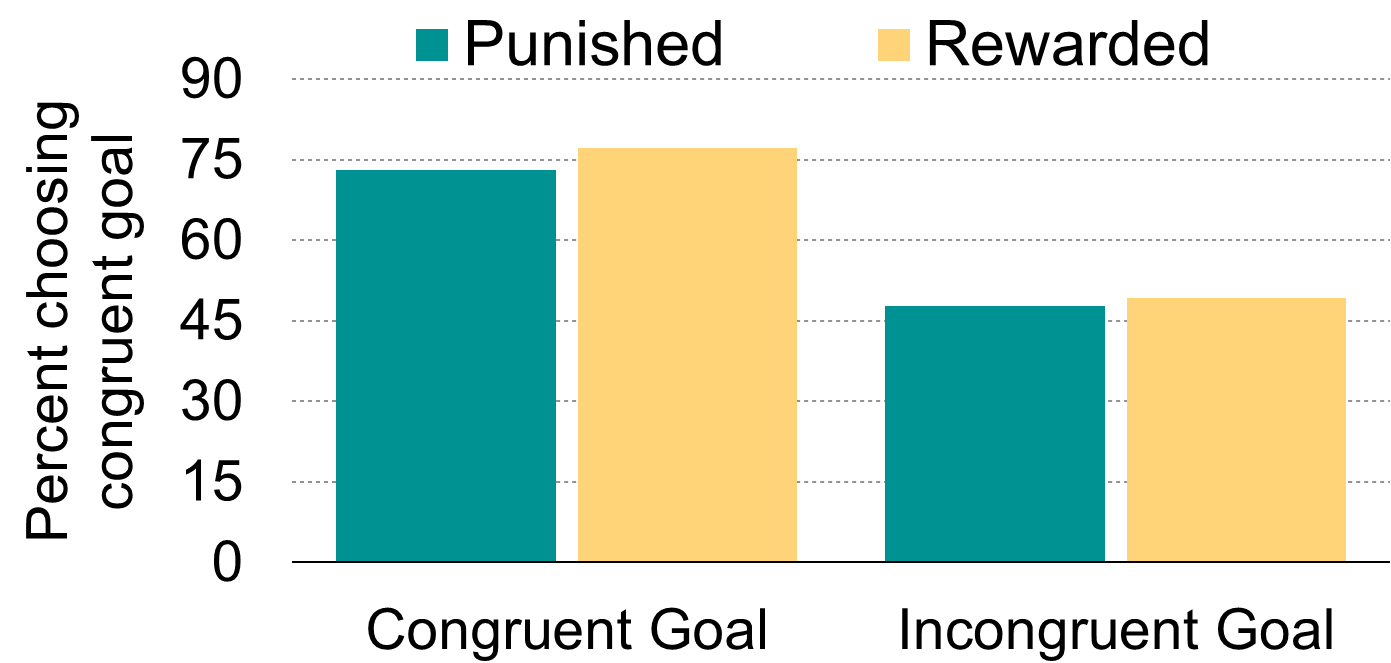
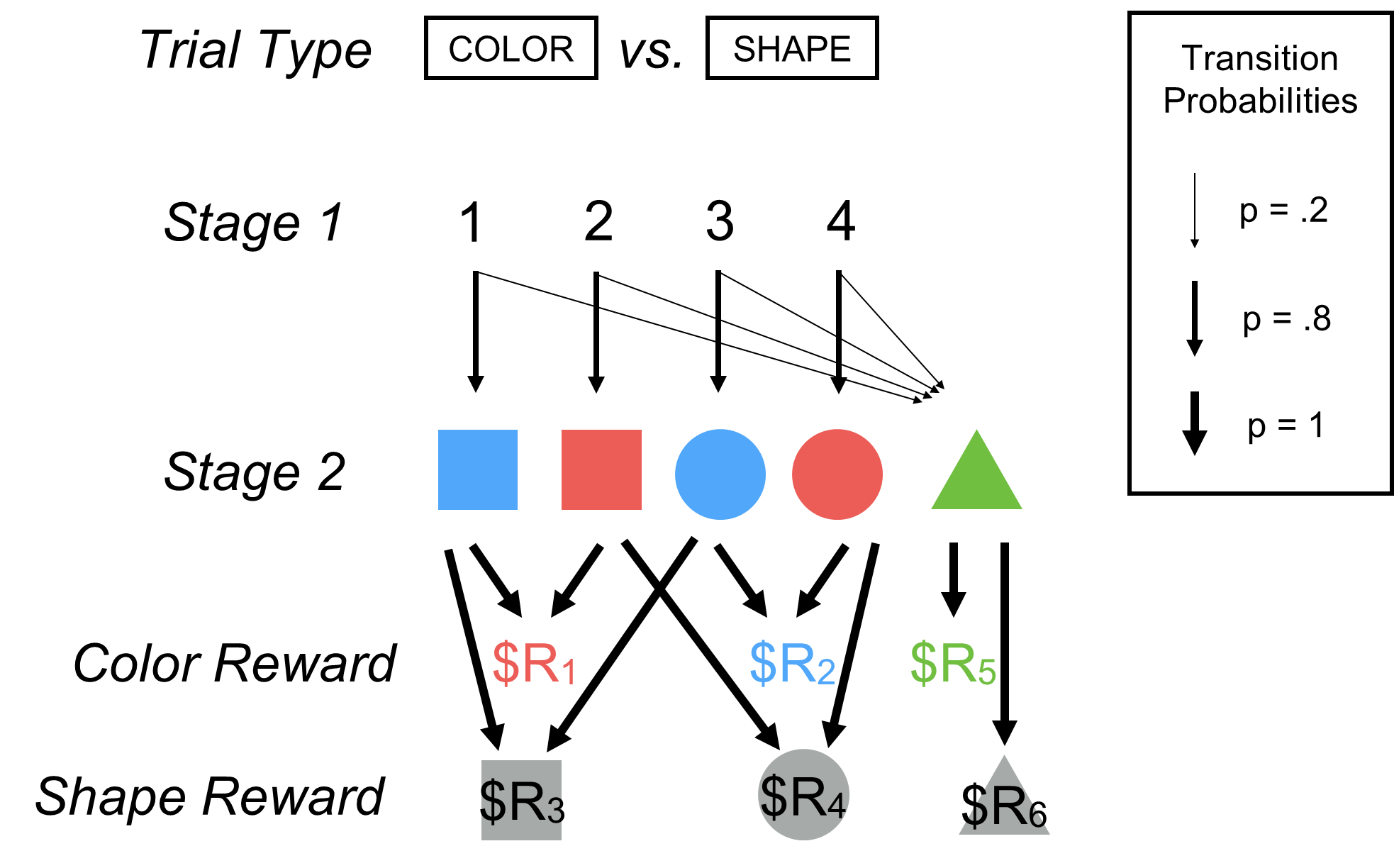
The results of Experiment 1 are consistent with a model of habitual control over goal selection, in which the selected goal then participates in a process of forward planning over a causal model of the transition structure of the task. They are also consistent, however, with hierarchical reinforcement learning models such as the options framework that do not invoke true planning17,21. We designed Experiment 2 to test whether the goals selected in our paradigm could be flexibly integrated with knowledge of independent state transitions, a hallmark of true planning.

Participants were first trained on a deterministic set of transitions between four Stage 0 options (A,B,C,D) and the same four Stage 1 options used in Experiment 1 (Figure 2A). During this initial training phase the Stage 1 options served as terminal states. Then, participants trained on and performed the task used in Experiment 1, without any involvement of Stage 0 choices. Finally, we tested each participant in a set of critical trials—those involving a low-probability transition followed by a goal-congruent choice—but presented participants with a pair of Stage 0 choices in place of the ordinary Stage 1 choices. In order to integrate information about a desired goal with the set of Stage 0 choices, participants were therefore required to engage in forward planning over the learned transition structure between Stage 0 and Stage 1. Here, again, we found that participants were significantly more likely to choose the congruent-goal option following positive reward (82%,SEM=1.0%) than following negative reward (68%,SEM=1.1%) *t*(292)=-10.9, *p*<.0001 (Figure 2B).



The statistical dependencies present in our task could support associative pairings between Stage 1 options22. For instance, participants might associate options 1 and 2 because they share a common high-probability transition to the blue state in Stage 2. Could such associative pairings account for the transfer of reward history from one option (e.g., 1) to another option (e.g., 2) on our critical trial type, and without the involvement of goal representations?

To investigate this possibility we conducted a third experiment (Figure 3A). Stage 2 states were arranged in a 2×2 design crossing color (red versus blue) and shape (circle versus square), with a fifth state that differed on both dimensions (a green triangle). Each trial was defined as a “color trial” or “shape trial”, with the trial type dictating the deterministic transitions to drifting rewards. Thus, three reward distributions were accessible on color trials, while three independent reward distributions were accessible on shape trials. Across trial types, each Stage 1 option was equally statistically associated to two other Stage 1 options: One by virtue of a shared Stage 2 color, and another by virtue of a Stage 2 shape. A model invoking associative pairings would predict equal transfer of reward history from any Stage 1 choice (e.g., A) to two other Stage 1 choices (e.g., B and C). Within a trial type, however, only a single Stage 2 dimension was goal relevant. A model-free influence on goal selection therefore predicts that reward history will influence the choice of goal-congruent options only for identical trial types: The reinforcement of a color goal would only influence color trials, and likewise for shape goals and shape trials. Our findings confirm this prediction. Following trials on which participants experienced a low probability transition to the green triangle state, they were more likely to choose the congruent goal Stage 1 option on the subsequent trial following reward (77%,SEM=1.2%) compared with punishment (73%,SEM=1.3%) *t*(386)=-3.3, *p*<.001. For incongruent goal trials, however, the effect was insignificant (positive: 49%,SEM=1.5%; negative: 48%,SEM=1.6%) *t*(360)=-.67, *p*=.51 (Figure 3B).



In sum, we find that goal selection is partially determined by model-free value representations derived from reward history. These goals are subsequently used during model-based planning according to an internally represented causal model of the task structure. Although our proposal depends upon the conceptual distinction between habitual (model-free) and goal-directed (model-based) behavioral control, it also posits a deep integration of these processes. This affords an explanation for empirical phenomena that blend canonical features of each process. For instance, many studies find that contextual cues can trigger goal pursuit outside of conscious awareness3, consistent with the operation of stimulus-response habits in the process of goal selection. There are also circumstances where goal-directed planning may be intrusive, such as *insert here*. Finally, it is commonly observed in educational settings that the execution of controlled cognitive processes improves with practice—in other words, that learning complex tasks requires the incremental acquisition of appropriate habits of thought [C]. It is widely recognized that human performance on such tasks involves hierarchically organized sets of goals and subgoals17,23. Proficient performance may thus entail the gradual acquisition of habitual subgoal selection given the contextual state of a superordinate goal [C].

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

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

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

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

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

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

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

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

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

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

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

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

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

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

15 Badre, D., Lebrecht, S., Pagliaccio, D., Long, N. M. & Scimeca, J. M. Ventral Striatum and the Evaluation of Memory Retrieval Strategies. (2014).

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

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

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

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

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

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

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

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