**Supplementary Materials**

**Experiment 2b**

???

**Regression analyses**

In order to capture trial-by-trial variation in the magnitude of the reward obtained on the setup trial, we regressed participants’ critical trial choices on the reward using a logistic mixed-effects model, estimating both random intercepts and random slopes at the subject level. (Following past research (CITE), 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 one 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.

In order 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).

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 SM Table 1 below. β is the coefficient of the reward regressor, χ2 is the statistic value in the likelihood ratio test, and p is the significance level of the likelihood ratio test.

|  |  |  |  |
| --- | --- | --- | --- |
| Experiment | β | χ2 | p |
| 1a | .15 | 309.93 | < .0001 |
| 1b | .032 | 17.7 | < .001 |
| 2a | .27 | 265 | < .0001 |
| 2b | .13 | 139.4 | < .0001 |

SM Table 1: Parameter estimates and significant tests for the mixed-effects models.

**Win-stay lose-shift analysis**

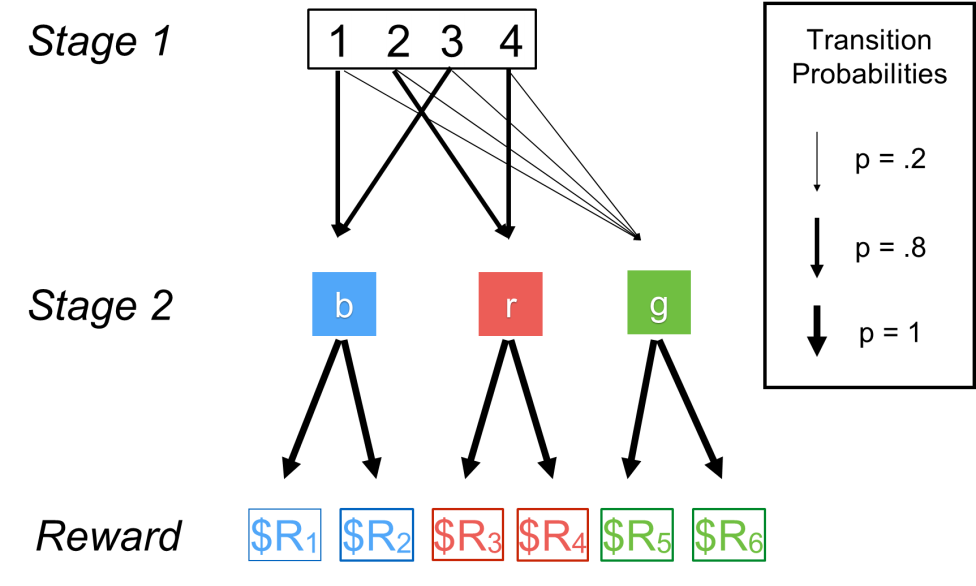
Our experiments demonstrated that setup trial reward influences goal choice on the next trial. These results are consistent with model-free valuation of goal choice, but they are also consistent with a win-stay lose-shift strategy over goal choice. In order to demonstrate that people are calculating true model-free goal values, we examined the influence of setup trial reward on participant choice *two* trials after the setup trial. A win-stay lose-shift strategy over goals would only directly impact participant choice on the first trial after reward. If there is additional influence of setup trial reward on the second trial, it must be due to model-free valuation of goal choice.

In order to control for any correlation between choices on the first and second trial, we separately examined cases in which the participant “stayed” and “shifted” on the trial immediately following the setup trial. In both cases, participants were significantly more likely in all experiments to choose shared-goal actions two trials after a reward than two trials after a punishment (all *p*s < .05). The sole exception was Experiment 1b, in which the effect was marginal for cases in which participants stayed on the intervening trial (*p* = .06), and nonexistent for cases in which the participants shifted (*p =* .42).

**Computational model**

We specified a computational model of learning and choice that includes model-free goal learning 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 model-based accuracy with model-free efficiency.

We generated simulated data for Experiment 1b. The task is a Markov decision process with ten states: The initial Stage 1 state, three Stage 2 states, and six reward states (SM Figure 1). The Stage 1 state had four possible actions, only two of which were available on any given trial. Each Stage 2 state had two available actions which led to reward states. The rewards were randomly generated for each agent by the same process as in the behavioral tasks, except we extended the reward boundaries from (-4,5) to (-8,8). This extension more sharply highlighted the contrast between the three mechanisms’ task performances.

Each agent completed 175 trials. Although agents made choices in both Stage 1 and 2, we focus exclusively on the Stage 1 choice because it effectively juxtaposes the three mechanisms.

SM Fig 1: The Markov decision task used in the simulations. Squares represent states and arrows represent actions.

*Mechanisms*

We implemented model-free goal learning with the options framework (CITE), a common framework for hierarchical policy abstraction. Under this framework, an “option” is a flexible policy which terminates upon attainment of a goal state. In our task, we defined two options available in Stage 1: one which terminates at blue (denoted *O1*), and the other at red (*O2*). (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 successfully 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,Oi)*. The values were initialized to zero and updated after every trial by:

where *Oi­* is the chosen option,*­ r* is the received reward, and *α* is a learning rate.

Agents used model-free updating to evaluate options, but model-based planning to achieve a chosen option’s goal state. Agents maintained the transition probabilities from each Stage 1 action *a* to each Stage 2 state *s*, denoted *T(a,s)*. Since participants were told these probabilities explicitly and had extensive practice with them, we assumed that agents knew the correct transition probabilities. (Our results do not change if agents learn the probabilities from experience.) Agents calculated option-specific action values by multiplying the probability that an actionwould reach the option’s goal state *gi* by a “pseudo-reward” *rpseudo* associated with obtaining *gi* (CITE). The value of choosing action *a* under option *Oi*, *U(Oi,a)*, was given by:

We set *rpseudo*= 1 for all goal states.

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

For comparison, agents also implemented a fully model-based hierarchical controller. The model-based controller was identical to the model-free goal learner except in its option evaluation mechanism. The model-based controller maintained values of each of the six reward states, denoted *V(sj)*, the transition probabilities from each Stage 2 action *a* to each reward state *sj*, denoted *T(a,sj)*, and the set of available actions in each Stage 2 goal state, denoted *A(gi)*.It calculated option value *V(Oi)* according to:

This model-based option evaluation mechanism is more accurate than our model-free mechanism because it disregards 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 *goal* learning, we implemented a traditional model-free action learner[[1]](#footnote-1). We used Q-learning, a common model of human learning and decision making (CITE). Agents maintained a value for each state-action pair, denoted *QMF(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 *α*:

We included eligibility traces, so the prediction error was applied to every previously chosen state-action pair utilized in that trial with decay parameter λ.

Since agents maintained three separate controllers with different state-action values, we produced a weighted mixture of the state-action values, *QW(s,a)*, by:

where *OMFG* and *OMB* are the options chosen by the model-free goal and model-based mechanisms respectively, and *wMFG* and *wMB* 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 *QW* values into a softmax function:

where *β* is a temperature parameter and *a1,2* are the two available actions in state *s*.

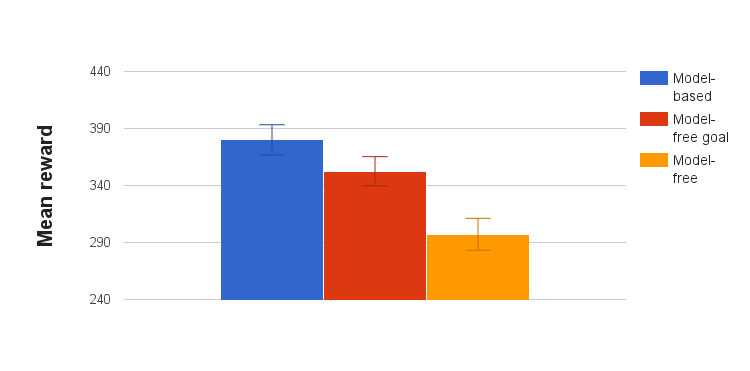
Thus, agents were characterized by five parameters: *α* (the learning rate), *λ* (the eligibility trace), *β* (the softmax temperature), *wMFG* (the model-free goal weight), and *wMB* (the model-based weight). Each agent’s parameters were randomly sampled as follows. *α* was sampled from a uniform distribution from 0 to 1, which we denote as U(0,1). *λ* was sampled from U(.5,1). *β* was sampled from U(.5,1.5). For the weights, three variables –  ,,and – were sampled from U(0,1), and then and . We generated 200 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 88.2% of the time after a reward and 79.2% of the time after a punishment (*t*(199) = -7.32, *p <* .0001). The mixed-effects model on same-type trials estimated a model-free goal coefficient of .072, and was preferred to a null model (χ2(2) = 101.7, *p* < .0001). In contrast, when agents did not perform model-free goal learning (), agents showed no difference in behavior following a reward versus a punishment (*t*(199) = -1.6, *p* > .1). Analysis by mixed effect models similarly showed null results (χ2(2) = 3.34, *p* > .1).

We also compared the performances of agents who exhibited only model-free goal selection (), model-based control (), or model-free control (). As predicted, our mechanism accumulated more total reward on the task than a pure model-free mechanism (*t*(199) = 5.4, *p* < .0001), but less than a pure model-based mechanism (*t*(199) = -3.3, *p* = .0012), suggesting that our mechanism balances the accuracy of model-based approaches with the computational efficiency of model-free approaches (SM Figure 2).

SM Figure 2: 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 380. A pure model-free mechanism, in yellow, earned 297. A model-free goal mechanism, in orange performed at an intermediate level, earning 353.



1. 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 our results through learned associations between shared-goal actions in the option-specific policies. However, by using novel action sets on critical trials, Experiments 2a and 2b rule out the possibility that our results could be entirely produced by a hierarchical model-free action controller. Therefore, we exclude this controller from our present analysis for ease of explanation. [↑](#footnote-ref-1)