**Supporting Information**

**Analysis of behavioral data by logistic regression**

In order 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.

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.

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

**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 ten states: The initial Stage 1 state *S1*, three Stage 2 states *S2-4*, and six reward states *S5-10* (Main Text Figure 1). At *S1* four possible actions exist, but only two of these were available on any given trial. *S2-4* 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 which terminates upon attainment of 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 *O1*), and the other with 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. After choosing option *Oi­* and transitioning to state *s’*,

where *r* is the received reward and *α* is a learning rate. We included eligibility traces, so the prediction error was applied to every previously chosen state-option pair in that trial with decay parameter λ.

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 *Sj*, denoted *T(S1,a,Sj)*. Since 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 intra-option 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 *Oi* according to:

where *A(S1)* is the set of actions available from *S1* on this trial, and *gi* ∈ {*S2-4*}is the goal state associated with option *Oi*.[[1]](#footnote-1)

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

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.[[2]](#footnote-2) To calculate the value of action *a1* in state *S1*,the model-based controller maintained the Stage 1 transition probabilities *T(S1,a1,Sj)*, the set of actions available from each Stage 2 state, denoted *A(Sj)*, the transition probabilities from each Stage 2 action *a2* ∈ *A(Sj)* to each reward state *Sk*,denoted *T(Sj,a2,Sk)*, and the current value of each reward state, denoted *V(Sk)*.(The values of reward states were learned by prediction error update after every trial with learning rate *α*.) It then calculated the value of option *a1* in state *S1* according to: [[3]](#footnote-3)

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.[[4]](#footnote-4) We used Q-learning, a common model of human learning and decision making (21). 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*:

As above, we included eligibility traces with the same 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 *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:

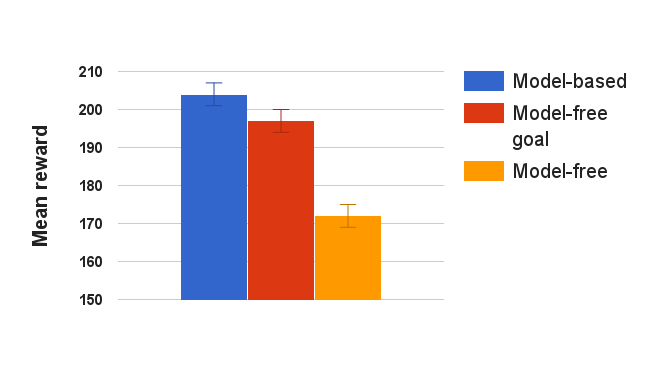
where *β* is a temperature parameter and *a1,2* are the two available actions in state *s*.

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 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% (± .7%) of the time after a reward and 66.7% (± .7%) of the time after a punishment. The mixed-effects model on same-type trials estimated a model-free goal coefficient of .1, and was preferred to a null model (χ2(2) = 343.1, *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 (71.3% vs. 71.5%). Analysis by mixed effect models similarly showed null results (χ2(2) = .422, *p* = .81).

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 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 (SI Figure 1).

*   
S. 1: 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.*

**Model fitting**

Our analysis of the effect of setup trial reward on critical trial choice, presented in the main text, suggests that participants spontaneously employ 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-R2s are presented in Table S2. The model fit the data significantly better than a chance model for every participant (likelihood ratio tests, all *p*s < .0001), and our parameter of interest, the model-free goal weight *wMFG*, was distributed significantly above zero (sign test, Z = 16.0, p < .0001).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Percentile** | *α* | *λ* | *β* | *wMFG* | *wMB* | -LL | Pseudo-R2 |
| 25th | .49 | .57 | .36 | .56 | .00 | 210 | .39 |
| Median | .65 | .97 | .49 | .74 | .00 | 185 | .46 |
| 75th | .85 | 1.00 | .62 | .96 | .11 | 158 | .54 |

*Table S2: Parameter estimates for participants in Experiment 1b. Shown are the 25th, 50th, and 75th percentiles of the distribution of each parameter across subjects. α is the learning rate, λ is the eligibility trace decay, β is the softmax temperature, wMFG is the relative weight of the model-free goal controller, and wMB is the relative weight of the model-based controller. Also shown are the distribution across subjects of negative log likelihoods and McFadden pseudo-R2 values (59), an approximate measure of the proportion of variance explained by the model.*

In order 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 *wMFG* 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 employing model-free control of goal selection.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | Aggregate -LL | Aggregate AIC | Exceedance probability | Number favoring full model |
| Full | 44223 | 90876 | 1.000 | -- |
| Without *wMFG* | 47194 | 96331 | 0.000 | 198 |

*Table S3: Model comparison between full model with model-free goal selection, and null model with wMFG set to zero. 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 < .05). 198 out of 243 subjects favored the full model.*

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 *wMFG* 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* = .89), and Bayesian model comparison indicated that the full model was heavily preferred to the null model (exceedance prob. = 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 prob. = 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.

1. Action probabilities under Stage 2 options were calculated similarly, using transitions from Stage 2 actions to terminating reward states. [↑](#footnote-ref-1)
2. 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. [↑](#footnote-ref-2)
3. Stage 2 action values were assigned by evaluating the inner sum of this equation. [↑](#footnote-ref-3)
4. 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 employ 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. [↑](#footnote-ref-4)