**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 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 |  |  |  |
| 1b |  |  |  |
| 2a |  |  |  |
| 2b |  |  |  |

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

We ran an additional regression analysis on the data from Experiments 2a and 2b. Our results are consistent with people maintaining a model-free valuation of the intermediate goals, but they are also consistent with people playing a win-stay lose-shift strategy (where “stay” and “shift” refer to intermediate goal choices). To demonstrate that people were choosing according to a valuation of the intermediate goals and not just win-stay lose-shift, we regressed critical trial choice on the value of the potentially *foregone* goal value. The foregone goal value was defined as the value of the goal which participants would forgo by staying with their previous goal choice on critical trials. For example, if a participant summed to 16 on a setup trial, then the foregone reward would be the value of summing to 21. (In keeping with our previous analysis, we estimated this value with the last reward obtained from choosing that goal. For example, if the last reward obtained from summing to 21 was +3, we would use +3 as the value of the foregone goal.)

If participants were playing win-stay lose-shift, then they would be insensitive to the value of the foregone goal. However, if they were making choices in proportion to model-free valuation of intermediate goals, then they would be sensitive to that value. We found that participants were indeed sensitive to value of the foregone goal. In Experiment 2, the value of the DO THIS???

*Simulations*

Parameters were sampled as follows.  was sampled from a uniform distribution from 0 to 1, which we denote as U(0,1).  was sampled from U(.5,1).  was sampled from U(0,1.5). For the weights, three variables –  ,,and – were sampled from U(0,1), and then  and .

**Computational model**

We specified a computational model of learning and choice that includes traditional model-based and model-free control alongside model-free goal learning. 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. (We simulate Experiment 2, which entails the key features of 1 and 3).

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

Agents behavior was determined by a weighted mixture of three controllers. Model-free control was implemented using a SARSA algorithm with eligibility traces (5). Agents estimated a model-free value of the state-action pair (*s,a*), denoted *MFV(s,a)*. In Stage 1, agents chose an action *a* and transitioned to state *s.* The value update for *MFV(1,a)* occurs by temporal difference learning with learning rate :



where *a’* is the only available action in state *s*.In Stage 2, agents chose *a’* and received reward *r*. Again, value update is given by temporal difference learning:



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



Model-based control was implemented a basic forward planning technique. Agents maintained a model-based value of each state-action pair, denoted *MBV(s,a)*. We assumed knowledge of the trial-type-dependent reward distributions on the part of the model-based controller. Thus, we separately indexed Stage 2 states according to the relevant trial type. To calculate the model-based value of each action from state 1, agents estimated the transition probability from *a* to *s*, denoted *T(a,s)*, by counting the number of observed transitions from *a* to *s* and normalizing the counters. Counters were initialized to 1, yielding a symmetric prior distribution over transition probabilities under the Dirichlet model. Then:



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



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



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



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





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

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

In the simulation with model-free goal learning, on same-type trials agents chose the shared-goal action 65.4% of the time after a reward and 55.5% of the time after a punishment (*t*(199) = -4.68, p < .0001). The simple mixed-effects model on same-type trials estimated a model-free goal coefficient of .063 (z = 5.47,p < .0001), and was preferred to a null model (*χ*2(2) = 40.9, p < .0001). The complete mixed-effects model showed similar results. In contrast, on different-type trials, agents chose what would have been the shared-goal action 49.3% of the time after a reward and 48.3% of the time after a punishment (*t*(191) = -.330, p = .75). Analysis by mixed effects models similarly showed null results for the model-free goal regressor (*p*s > .8).

In the simulation where agents did not perform model-free goal learning agents showed no difference in behavior following a reward versus a punishment on same-type trials (*t*(199) = 1.30, p = .195). Analysis by mixed effect models similarly showed null results (*p*s > .6).