**Supplementary Materials**

**Experiment 1**

*Methods*

232 subjects were recruited on Amazon Mechanical Turk to participate in a two-stage task. Stage 1 had four options (represented by the numbers 1-4), which each led probabilistically to one of three states (represented by the colors red, blue, and green). These states in turn had only one available action, which deterministically led to a reward. The task’s structure is depicted schematically in Figure 1.

*INSERT FIGURE 1*

1 and 3 each had an 80% chance of going to blue, and a 20% chance of going to green. 2 and 4 each had an 80% chance of going to red, and a 20% chance of going to green. The high-probability transition of each number became the “goal” of that number – the goal of clicking on 1 would be to get blue, the goal of clicking on 2 would be to get red, etc. Subjects were explicitly told these probabilities, and were trained on them in the practice rounds.

In each trial, subjects were first presented with two of the four numbered options. The presented options were chosen randomly, with two constraints. The first constraint was that the high-probability transitions of the two options had to lead to different colors. (Thus, the possible option pairs were: 1 and 2, 1 and 4, 2 and 3, 3 and 4.) The second constraint was that the selected action on a pre-critical trial could not be an available option on a critical trial[[1]](#footnote-1). (So if the options on the pre-critical trial were 1 and 2, and the subject chose 1, the only possible option pairs for the critical trial would be 2 and 3 or 3 and 4.)

After clicking on one of the two numbers, subjects were shown the color they received. The transitions from number to color were determined randomly according to the transition probabilities above – except in the case of pre-critical trials, which always transitioned to green. (The transition probabilities were adjusted such that, factoring in the predetermined amount of critical trials, the subject would see high-probability transitions on 80% of all trials and low-probability transitions on 20% of all trials.)

Once subjects had received a color, they clicked on the color and received a reward. Each color had an associated reward. The rewards for each color were initialized uniformly at random on a range of -4 points to +5 points, and went on a bounded Gaussian random walk for the rest of the experiment. After each round, the drift was sampled from a normal distribution with (μ=0, σ=1.8), rounded to the nearest integer, and added to the current reward level[[2]](#footnote-2). On pre-critical trials, the reward was temporarily boosted to the extremes of the reward range. If the underlying reward level for green was positive, then, on the pre-critical trial only, the subject would receive: Similarly, if the reward level for green was negative, the subject would receive: . One point corresponded to one cent of bonus money.

Subjects had 75 practice trials and 175 real trials. On the real trials, subjects had only 4 seconds to make their choice between the two numbers. If they did not make a choice within 4s, the trial would time out and the next trial would begin. 14 subjects were excluded from analysis because they timed out on more than 50 trials, leaving 218 subjects for all the analyses.

*Analyses*

We restricted our primary analyses to critical trials. The trials before critical trials, called the pre-critical trials, were automatically low-probability transitions to green. Then, on the critical trials themselves, the action selected in the pre-critical trial was not available as an option in the critical trial. These two manipulations ensured that neither a pure model-based nor a pure model-free system could learn from the pre-critical trial reward in a way that would affect the choice on the critical trial. The only system which could incorporate that reward would be a system of model-free goal selection. See Figure 2 for an example of a critical trial.

*INSERT FIGURE 2*

The independent variable in all our analyses was the pre-critical trial reward (called the *model-free-goal* value). The dependent variable was the subject’s choice on the subsequent critical trial, coded as 1 if the chosen action had the same goal as the previous round and 0 otherwise. In the above example, the model-free-goal value would be +??, and the dependent variable would be 1 if the subject chose ??, and 0 otherwise.

We hypothesized that the model-free-goal value would predict subjects’ choices on the subsequent critical trials. We tested this hypothesis in three ways. First, as a coarse test, we computed the mean choice when the model-free-goal value was less than 0 and the mean choice when it was greater than 0, and compared the two using a paired t-test. Second, as a more granular test, we regressed choice on reward using a logistic mixed-effects model (with both a random intercept and slope).

Third, to definitively rule out any influence from a pure model-based or model-free system, we estimated a second mixed-effects model with approximate model-based and model-free action values as additional regressors. These values were calculated as follows. The model-based value of an action with a certain color goal was defined as the last reward which the subject received from that color, discounted over time (with discount parameter ). If the last time a subject got to the color blue was *t1* rounds ago, and she received *r1* points, the model-based value of numbers 1 and 3 would be . On the other hand, the model-free value of an action was defined as the reward received the last time the subject selected that action, discounted over time. If the last time a subject selected number 3 was *t2* rounds ago, and she received *r2* points, the model-based value of 3 would be . was fixed at .85.

These two values, model-based and model-free, were computed for both available action options in each critical trial. Then the model-based value of the action which the subject did not chose was subtracted from the model-based value of the action which the subject did chose (in accordance with the coding scheme of the dependent variable), and the resulting single value became the model-based regressor in the mixed-effects model. The same procedure was applied to the model-free values. Therefore, the second mixed-effect model had three regressors: model-based, model-free, and model-free-goal. The model estimated both fixed and random slopes for all three regressors[[3]](#footnote-3).

To test the significance of the model-free-goal regressor in the mixed-effect models, we estimated null models (which in each case was the full model with the model-free-goal regressor removed), and performed both likelihood ratio tests and parametric bootstrap analyses to assess whether the model-free-goal regressor increased the model’s likelihood enough to be justifiably included.

Finally, to benchmark the size of the model-free-goal effect, we also considered trials which followed low-probability transitions to green but were not critical trials, because the option chosen in the trial with the low-probability transition was available again. We call these *congruent action* trials. See Figure 3 for an example.

*INSERT FIGURE 3*

On congruent action trials, the effect of the reinforcement on the low-probability transition could be incorporated by both a model-free and a model-free-goal system. Therefore, subtracting the effect size of the reinforcement on the congruent goal trials from the effect size on these trials could give us a rough estimate of the size of the model-free effect alone. We could then compare that magnitude to the magnitude of the model-free-goal effect. To measure the effect size of reinforcement on congruent action trials, we ran both the paired t-test and the simpler mixed-effect model on them.

*Results*

There were a total of 6120 congruent goal trials over 218 subjects. When the model-free-goal was positive, subjects chose the congruent goal 85.2% of the time. When it was negative, subjects chose the congruent goal 68.7% of the time. This difference was significant (paired t-test, t(217) = -11.2, p < .0001).

In the simple mixed-effect model, with only the model-free-goal reward as a regressor, the coefficient of the reward was .152 (Wald test, z = 11.5, p < .0001). The model was preferred to a null model without the reward (Likelihood ratio test, (2) = 312.1, p < .0001). In a parametric bootstrap analysis, ??? out of 1000 randomly resampled null models had a likelihood as large as the full model (???).

In the complete mixed-effect model, with the model-based, model-free, and model-free-goal values as regressors, the coefficient of the model-free-goal value was .157 (z = 11.5, p < .0001). The model was preferred to the null model ((4) = 326.6, p < .0001). In a parametric bootstrap analysis, ??? out of 1000 randomly resampled null models had a likelihood as large as the full model (p < ???). For comparison, the model-based coefficient was .152 (z = 6.37, p < .0001) and the model-free coefficient was .071 (z = 3.26, p < .005).

There were a total of 539 congruent action trials. On these trials, after a reward subjects chose the same action 83.4% of the time. After a punishment, subjects chose the same action 72.3% of the time. This difference was significant (t(117) = -2.47, p < .02), and the effect size was ???[[4]](#footnote-4). In the simple mixed-effect model, the coefficient for model-free-goal was .188. ???

**Experiment 2**

*Methods*

312 subjects were recruited online through Amazon Mechanical Turk. They performed a task identical to the one above, with one change. Before being exposed to the structure of the main task, subjects were trained on a set of intuitive, deterministic transitions from letters to numbers. Those transitions are depicted in Figure 3.

*INSERT FIGURE 3*

After becoming familiar with those transitions, subjects proceeded with the same task as above. All non-critical trials had exactly the same structure as above, with a choice between two numbers leading to a color, which in turn led to one of three drifting reward distributions. However, on critical trials, subjects instead were presented with a choice between two letters. Subjects chose a letter and received a number (in accordance with the deterministic transitions in Figure 3). They then clicked on that number and, in the usual way, got to a color which led them to a reward. The critical trials thus required a goal-directed system to plan one extra step ahead. The structure of critical trials versus non-critical trials is depicted in Figure 4.

*INSERT FIGURE 4*

*Analysis*

All analyses were identical to those in Experiment 1.

*Results*

There were a total of 8086 congruent goal trials over 293 subjects. When the model-free-goal reward was positive, subjects chose the congruent goal 81.9% of the time. When it was negative, subjects chose the congruent goal 67.7% of the time. This difference was significant (t(292) = -10.9, p < .0001).

In the simple mixed-effect model, the model-free-goal coefficient was .118 (z = 11.3, p < .0001). The model was preferred to a null model ((2) = 291.6, p < .0001). In a parametric bootstrap analysis, ??? out of 1000 randomly resampled null models had a likelihood as large as the full model (p < ???).

In the complete mixed-effect model, the model-free-goal coefficient was .118 (z = 11.3, p < .0001). The model was preferred to a null model ((4) = 291.6, p < .0001). In a parametric bootstrap analysis, ??? out of 1000 randomly resampled null models had a likelihood as large as the full model (p < ???). For comparison, the model-based coefficient was .125 (z = 6.92, p < .0001) and the model-free coefficient was .055 (z = 2.92, p < .005).

There were a total of 677 congruent action trials. On these trials, after a reward subjects chose the same action 84.8% of the time. After a punishment, subjects chose the same action 70.2% of the time. This difference was significant (t(117) = -3.68, p < .001), and the effect size was ???. In the simple mixed-effect model, the coefficient for model-free-goal was .197. ???

**Experiment 3**

*Methods*

191 subjects were recruited online through Amazon Mechanical Turk. They performed a task identical to the one in Experiment 1, with the following change. In Experiment 1, stage-2 states only varied in their color (blue, red, or green). In Experiment 3, they also varied in their shape. There were three shapes: square, circle, and triangle. The new objects and transition probabilities are depicted in Figure 5(a).

*INSERT FIGURE 5*

In Experiment 1, the color of an object determined the reward associated with it. On any given trial, blue had a certain reward, red had a certain reward, and green had a certain reward. In Experiment 3, the reward value of an object could either be determined by the color or shape of the object. Each color and shape had a separate drifting reward distribution associated with it.

There were two trial types. On color trials, it was the color of the object which determined the reward. On shape trials, it was the shape of the object. Before each trial began, subjects were told the trial type. The flow of Experiment 3 is depicted in Figure 5(b).

In Experiment 3, there were also two types of critical trials. In congruent critical trials, the trial type of the critical trial was the same as the trial type of the pre-critical trial. In incongruent critical trials, the two trial types were different.

*Analyses*

Our hypothesis only holds for congruent critical trials, not incongruent critical trials. Therefore, the same analyses as in Experiments 1 and 2 were conducted on the congruent critical trials to show that the model-free-goal value predicted subjects’ choices.

In addition, we wanted to show that the effect was selective to congruent critical trials. To that end, we conducted two additional analyses. First, we conducted all three tests (paired t-test and two mixed-effect models) on incongruent critical trials to see whether the effect was present. Second, we estimated a mixed-effect model with both congruent and incongruent critical trials, where we included the type of critical trial as a regressor interacting with the model-free-goal regressor[[5]](#footnote-5).

*Results*

*INSERT RESULTS HERE*

**Simulations**

*Methods*

To validate our experiments, we built a generative model of agents who used some combination of model-based, model-free, and model-free-goal learning. We then ran the agents through the same game as in our experiments, and showed that our experiment detected a result if and only if the agents were actually using model-free-goal learning.

The game was implemented as a Markov decision process with six states, the initial stage-1 state and then one state for each stage-2 object. State 1 had four possible actions (i.e. the four numbers), only two of which were available on any given trial. States 2-6 had only one possible action (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.

The agents had three learning mechanisms. Their model-free reinforcement learning mechanism was the SARSA algorithm with eligibility traces (cite). 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.* Their model-free update was:

In stage 2, agents chose the only available action (i.e. clicking on the object) and received reward *r*. Their model-free updates were:

was the learning rate, and was the eligibility trace. Parameter selection is described below.

Agents’ model-based learning mechanism was ???. Agents maintained a model-based value of each state-action pair, denoted *MVB(s,a)*, but the model-based mechanism had a different conception of states. To the model-free system, each different object was a different state. But because the model-based system knew that rewards were tied to object features (color or shape), not objects themselves, it conceptualized each object *feature* as a different state. The feature-state corresponding to a given object-state *s* is given by *.*

To calculate the model-based value of each action from state 1, agents estimated the transition probabilities of action *a* to state *s*, denoted by , by dividing the number of observed transitions from *a* to *s* by the total number of times *a* was selected. Then:

After transitioning to state *s*, clicking on the color, and receiving reward *r*, the model-based update was:

The third learning mechanism was our proposed mechanism, 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 calculated a model-free value of each possible feature goal, denoted *MFG(g)*, by:

To determine the probability of an action *a*, agents took a weighted average *Wa* of the three values and entered it into a softmax function:

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

We ran two simulations, one with only color trials (corresponding to Experiment 1) and one with both color and shape trials (corresponding to Experiment 3). For each simulation, 200 agents were generated with randomly sampled parameters[[6]](#footnote-6). They played through the game, and their results were analyzed by the same process as in the behavioral tasks.

*Results*

1. For a more detailed treatment of critical trials, see the next section. [↑](#footnote-ref-1)
2. If a drift would have caused a reward level to exceed +5 or -4, the reward would ‘rebound’ by however much it would have gone over. For example, if the reward were at +3 and the drift were +5, the next reward would be . [↑](#footnote-ref-2)
3. In all our mixed-effects models, we allowed full correlation among all the random slopes and the random intercept. [↑](#footnote-ref-3)
4. There are fewer degrees of freedom in this t-test because only 118 subjects saw any congruent action trials. [↑](#footnote-ref-4)
5. We first estimated a model with the two main effects (type of critical trial and model-free-goal value) and an interaction, but the model was overspecified. Dropping the main effect of critical trial type eliminated that overspecification. Thus, we restrict our presented results to the model with the main effect of model-free-goal value and the interaction between model-free-goal value and type of critical trial. [↑](#footnote-ref-5)
6. was sampled from a uniform distribution from 0 to 1, which we will 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 . [↑](#footnote-ref-6)