Supporting Information

"Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems"

Reinforcement learning model

We fitted choices on the two-step tasks to an established and validated dual-system reinforcement-learning model (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Daw, Niv, & Dayan, 2005; Gläscher, Daw, Dayan, & O'Doherty, 2010). The original version of the two-step task consists of three states across two stages, both with two available actions (a_A and a_B), whereas our novel paradigm consists of four states across two stages, with two available actions at the first-stage states (a_A and a_B) and one action at the second-stage state (a_C). Our models consist of model-based and model-free strategies that both learn a function $Q(s, a)$ mapping each state-action pair to its expected future return (value). On trial t, the first-stage state is denoted by $s_{1,t}$, the second-stage state by $s_{2,t}$, the first- and second-stage actions by $a_{1,t}$ and $a_{2,t}$, and the second-stage rewards as $r_{1,t}$ (always zero, there is only reward on the second stage) and $r_{2,t}$.

Model-free strategy. The model-free agent uses the SARSA(λ) temporal difference learning algorithm (Rummery & Niranjan, 1994), which updates the value for each state-action pair (s, a) at stage i and trial t according to:

$$Q_{MF}(s_{i,t}, a_{i,t}) = Q_{MF}(s_{i.t}, a_{i,t}) + \alpha\delta_{i,t}e_{i,t}(s, a)$$

where

$$\delta_{i,t} = r_{i,t} + Q_{MF}(s_{i+1,t}, a_{i+1,t}) - Q_{MF}(s_{i,t}, a_{i,t})$$

is the reward prediction error, α is the learning rate parameter (which determines to what degree new information is incorporated), and $e_{i,t}(s, a)$ is an eligibility trace set equal to 0 at the beginning of each trial and updated according to

$$e_{i,t}(s_{i,t}, a_{i,t}) = e_{i-1,t}(s_{i,t}, a_{i,t}) + 1$$

before the Q-value update. The eligibilities of all state-action pairs are then decayed by λ after the update.

We now describe how these learning rules apply specifically to the two-step task. The reward prediction error is different for the first two stages of the task. Since $r_{1,t}$ is always zero, the reward prediction error at the first stage is driven by the value of the selected second-stage action $Q_{MF}(s_{2,t}, a_{2,t})$:

$$\delta_{1,t} = Q_{MF}(s_{2,t}, a_{2,t}) - Q_{MF}(s_{1,t}, a_{1,t})$$

Since there is no third stage, the second-stage prediction error is driven by the reward $r_{2,t}$:

$$\delta_{2,t} = r_{2,t} - Q_{MF}(s_{2,t}, a_{2,t})$$

Both the first- and second-stage values are updated at the second stage, with the first-stage values receiving a prediction error down-weighted by the eligibility trace decay, λ. Thus, when $\lambda = 0$, only the values of the current state get updated.

Model-based strategy. The model-based algorithm works by learning a transition function that maps the first-stage state-action pairs to a probability distribution over the subsequent states, and then combining this function with the second-stage model-free values (i.e., the immediate reward predictions) to compute cumulative state-action values by iterative expectation. In other words, the agent first decides which first-stage action leads to which second-stage state, and then learns the reward values for the second-stage actions.

At the second stage, the learning of the immediate rewards is equivalent to the model-free learning, since those Q-values are simply an estimate of the immediate reward $r_{2,t}$. As we showed above, the SARSA learning rule reduces to a delta-rule for predicting the immediate reward. This means that the two approaches coincide at the second stage, and so we set $Q_{MB} = Q_{MF}$ at this stage.

The model-based values are defined in terms of Bellman's equation (Sutton & Barto, 1998), which specifies the expected values of each first-stage action using the transition structure P (assumed to be fully known to the agent):

$$Q_{MB}(s_A, a_j) = P(s_B|s_A, a_j) \max_{a\in\{a_A,a_B\}} Q_{MF}(s_B, a) + P(s_C|s_A, a_j) \max_{a\in\{a_A,a_B\}} Q_{MF}(s_C, a)$$

where we have assumed these are recomputed at each trial from the current estimates of the transition probabilities and second-stage reward values.

Decision rule. To connect the values to choices, the Q-values are mixed according to a weighting parameter w:

$$Q_{net}(s_A, a_j) = wQ_{MB}(s_A, a_j) + (1 - w)Q_{MF}(s_A, a_j).$$

To accommodate our stake manipulations, we defined two different weights that operated on different trial types. We set $w = w_{\text{low}}$ on trials with low stakes, and $w = w_{\text{high}}$ on high stakes trials.

For the model in Experiment 2, at the second stage the decision is made using only the model-free values, whereas there was no choice at the second stage in Experiment 1. We used the softmax rule to translate these Q-values to actions. This rule computes the probability for an action, reflecting the combination of the model-based and model-free action values weighted by an inverse temperature parameter. At both states, the probability of choosing action *a* on trial *t* is computed as

$$P(a_{i,t} = a|s_{i,t}) = \frac{\exp\left(\beta\left[Q_{net}(s_{i,t}, a) + \pi \cdot rep(a) + \rho \cdot resp(a)\right]\right)}{\sum_{a'} \exp\left(\beta\left[Q_{net}(s_{i,t}, a') + \pi \cdot rep(a') + \rho \cdot resp(a')\right]\right)}$$

where the inverse temperature β determines the randomness of the choice. Specifically, when $\beta \to \infty$ the probability of the action with the highest expected value tends to 1, whereas for $\beta \to 0$ the probabilities over actions becomes uniform. The indicator variable rep(*a*) is defined as 1 if *a* is a first-stage action and is the same one as was chosen on the previous trial, zero otherwise. Multiplied with the 'stickiness' parameter π, this captures the degree to which participants show perseveration ($\pi > 0$) or switching ($\pi < 0$) at the first stage. The indicator variable resp(*a*) is defined as 1 if *a* is a first-stage action selecting the same response key as the key that was pressed on the previous trial, zero otherwise. Multiplied with the 'response stickiness' parameter ϱ, this captures the degree to which participants repeated ($\varrho > 0$) or alternated ($\varrho < 0$) key presses at the first stage. We introduced this parameter since the spaceships' positions were not fixed, hence participants could show perseveration in spaceship choices, button presses, or both.

Model fitting procedure

We used maximum *a posteriori* estimation with empirical priors, implemented using the *mfit* toolbox (Gershman, 2016) parameters to fit the free parameters in the computational models to observed data. Based on prior work (Gershman, 2016), we used weak priors for the distributions for the inverse temperature, $\beta \sim$ Gamma(4.82, 0.88), and stickiness parameters, $\pi, \varrho \sim \mathcal{N}(0.15, 1.42)$, and flat priors for all other parameters. To avoid local optima in the estimation solution, we ran the optimization 100 times for each participant with randomly selected initializations for each parameter. The final estimations for β, *a*, π, ϱ, and *w*, were extracted from the run with the maximal log-likelihood and are reported in Table 1.

Exhaustive reinforcement-learning model

It is possible that the difference between of weighting parameters of the two stake-size conditions was affected by changes in behavior that were unrelated to a difference in allocation between model-based and model-free reinforcement-learning strategies. In the model described above, we only varied the weighting parameter, possibly forcing other behavioral changes caused by the stake manipulation only on this parameter. To address this concern, we developed a version of the reinforcement-learning model that varies all parameters between the high- and low-stake conditions.

Specifically, depending on the trial's stake size, this exhaustive model made its choices between actions and updated their values using either for β_{low}, a_{low}, π_{low}, ϱ_{low}, and w_{low}, or β_{high}, a_{high}, π_{high}, ϱ_{high}, and w_{high}. Everything else was identical to the dual-system reinforcement-learning model described above. For both experiments, we used the same maximum *a posteriori* estimation with empirical priors to obtain estimates for these 12 parameters.

Results

The results of these analyses are given in Table S1. The inverse temperature parameter β was affected by the stake size for both Experiment 1 [$t(97) = 3.95$, $p < 0.001$, Cohen's $d = 0.40$], and Experiment 2 [$t(99) = 4.00$, $p < 0.001$, Cohen's $d = 0.40$]. For both studies, this effect indicated that participants showed more exploiting behavior when the stakes were high. Replicating our previous findings we still obtained a significant effect of stake size on the weighting parameter in Experiment 1 [$t(97) = 3.15$, $p = 0.002$, Cohen's $d = 0.32$], but not for Experiment 2 [$t(99) = 0.48$, $p = 0.63$, Cohen's $d = 0.05$]. The difference in these effects also reached statistical significance, $t(196) = 2.47$, $p = 0.014$, Cohen's $d = 0.35$. The remaining parameters in either experiment did not significantly differ between the high- and low-stake conditions for Experiment 1, ps > 0.10, and Experiment 2, ps > 0.50. This pattern of results suggests that the increase in the weighting parameter on high stakes trials in Experiment 1 cannot be fully explained by changes in behavior unrelated to the difference in reinforcement-learning strategies. They also indicate that the lack of an effect in Experiment 2 was not simply caused by participants' lack of attention to the stake cues, since they still showed an increase in exploitation behavior when the stakes were high.

Table S1. Best-fitting parameter estimates of the exhaustive model shown as median plus quartiles across participants for both experiments.

	Percentile	β_{low}	β_{high}	a_{low}	a_{high}	λ_{low}	λ_{high}	π_{low}	π_{high}	ϱ_{low}	ϱ_{high}	w_{low}	w_{high}
Exp. 1	25th	0.49	0.68	0.01	0.01	0.00	0.00	-0.12	-0.07	-0.31	-0.27	0.00	0.48
	Median	0.79	1.04	0.70	0.69	0.49	0.3	0.08	0.08	-0.11	-0.12	0.65	0.87
	75th	3.25	3.38	1.00	1.00	1.00	0.92	0.61	0.46	0.1	0.04	0.97	1.00
Exp. 2	25th	2.50	3.00	0.01	0.00	0.30	0.40	0.01	0.04	-0.02	-0.02	0.00	0.00
	Median	3.35	3.57	0.15	0.16	0.66	0.60	0.17	0.15	0.05	0.03	0.17	0.31
	75th	3.82	4.34	0.47	0.48	1.00	1.00	0.32	0.30	0.15	0.15	0.83	0.66

Multilevel logistic regression analysis

In addition to the model-fitting procedure described above, we also investigated choice behavior on this task by analyzing the probability of repeating choices from trial to trial using multilevel logistic regression models. These analyses were carried out with Matlab's *fitlme* function.

Experiment 1

In this paradigm of Experiment 1, the implicit equivalence between the two first-stage states allows for a dissociation between habitual and goal-directed choice (Doll et al., 2015) based on the probability with which participants repeat the second-stage state. The model-based strategy uses the experiment's structure to plan towards the second-stage model-free values, allowing it to generalize knowledge learned from both starting states. Thus, outcomes at the second stage equally affect first-stage preferences, regardless of whether this trial starts with the same starting state as the previous trial. This contribution is reflected in a main effect of the prediction error sign on stay probability, since the model-based strategy is insensitive to changes in starting state. For the model-free strategy, however, rewards that are received following one start state should not affect subsequent choices from the other start state. The model-free learner only shows increased stay probability when the current start state is the same as that on the previous trial, and this is reflected as an interaction between previous outcome and starting state.

To test our cost-benefit hypothesis, we used a multilevel logistic regression analysis to investigate whether the stake manipulation affected the strength of these two effects. This model predicted whether participants repeated the previous trial's second-stage state (i.e., "staying behavior") as a function of the similarity of the previous trial's first-stage state, the previous trial's prediction error sign (estimated using the computational model and individual parameter fits), and the stake condition. Specifically, the dependent variable was whether the current second-stage choice was the same as that on the previous trial. For each trial, the predictors for this analysis were the sign of the prediction error on the previous trial on the previous trial (RPE_{i-1}), whether the previous starting state was the same or different from the current starting state ($same_i$), and the size of the stake on the current trial ($stake_i$). The final multilevel regression model included these three predictors, their interactions and the intercept. We modeled all coefficients as random effects, varying between participants around a group mean.

In this regression analysis, the main effect of RPE_{i-1} represents the model-based contribution, since it carries over to the next trial even when the start states are different, whereas the interaction term $RPE_{i-1} \times same_i$ captures effects that are specific to the state in which they were received and therefore represent the model-free contribution.

The results from this regression analysis are given in Table S2. We found significant effects of the regressors for the main effect of the sign of the previous trial's prediction error, indicating a model-based contribution, and the interaction effect between previous prediction error and the similarity of the current and previous first-stage states, indicating the model-free contribution, ps < 0.001. Importantly, we found that the model-based effect, i.e., the previous trial's prediction error's main effect, was significantly stronger on high stake trials, [$t(17919) = 4.10$, $p < 0.001$]. The model-free effect, the strength of the interaction between starting similarity and previous prediction error sign, was not modulated by the stakes manipulation, [$t(17919) < 1$].

Table S2. Regression coefficients from multilevel logistic regression analysis for Experiment 1, indicating the effect of outcome of previous trial, similarity of previous starting state to current starting state, and stake condition, on repetition of second-stage choice.

Coefficient	Estimate (SE)	p
(Intercept)	.64 (.07)	< .001
Previous RPE	.31 (.03)	< .001
Same starting state	.11 (.03)	< .001
Stake size condition	.03 (.01)	< .001
RPE × Same	.07 (.02)	< .005
RPE × Stake size	.03 (.01)	< .001
Same × Stake size	.01 (.01)	.30
RPE × Same × Stake size	-.00 (.01)	.94

The finding that the model-free component was not affected by stake size seems inconsistent with a model in which the systems are in direct competition. However, a subsequent series of multilevel logistic regression

analyses on 100 simulated data sets using the median fits and the computational model reported in Table 1 show that the model-based component was affected by stake size in 99% of simulations, but the model-free component in only 9% of simulations. This shows that these logistic regression analyses have differential sensitivity in estimating the strength of the two effects, and therefore that the logistic regression result does not contradict our computational model.

Experiment 2

In addition to computation model analysis, we again investigated behavior on this task using a multilevel logistic regression analysis.

Here, the dissociation between model-based and model-free control in our novel paradigm follows a different logic than in Experiment 1. Since the model-free strategy is insensitive to the task structure, it will simply increase the probability of staying with the previous action if it led to reward, regardless of the type of transition on the previous trial. This contribution is reflected as a main effect of previous outcome on the stay probability. The model-based strategy is reflected by an interaction between previous transition type and outcome, because it decreases the stay probability after a reward and a rare transition in order to achieve a higher likelihood to get to the previously rewarded second-stage state. After a rare transition and a loss, the model-based strategy is more likely to stick with the original action, since this decreases the likelihood of getting to the unrewarded state.

We again used a multilevel logistic regression analysis to investigate whether the stake manipulation affected the strength of the model-based and model-free components. This model predicted whether the current trial's first stage choice was the same the previous trial's first-stage choice state (i.e., "staying behavior") as a function of whether the previous trial produced a reward (r_{i-1}), what type of transition occurred on that trial ($common_i$), and the size of the stake on the current trial ($stake_i$). The multilevel regression model included these three predictors, their interactions and the intercept. All these coefficients were included as random effects, varying between participants around a group mean.

In this regression analysis, the main effect of r_{i-1} represents the model-free contribution, since it captures reward effects that are independent from the transition type on the previous trial, whereas the interaction term $r_{i-1} \times common_i$ captures reward effects that are modulated by the transition type on the previous trial and therefore represents the model-based contribution.

The results from the logistic regression for the Daw paradigm are given in Table S3. The regressor for the main effect of the outcome of the previous trial was significant, $p < 0.001$, indicating a model-free contribution. The regressor for the interaction between previous outcome and previous transition type was significant as well, $p < 0.001$, indicating a model-free contribution. Most importantly, we found that that neither model-free effect (i.e., the main effect of the previous outcome), nor the model-based effect (the interaction between the previous outcome and previous transition type) were significantly affected by the stake size manipulation, [$t(18393) = 1.37$, $p = 0.17$] and [$t(18393) < 1$], respectively.

Table S3. Regression coefficients from multilevel logistic regression analysis for Experiment 1, indicating the effect of outcome of previous trial, similarity of previous starting state to current starting state, and stake condition, on repetition of second-stage choice.

Coefficient	Estimate (SE)	*p*
(Intercept)	.83 (.09)	< .001
Previous outcome	.22 (.04)	< .001
Previous transition	.04 (.02)	0.06
Stake size	.02 (.02)	.25
Outcome × Transition	.16 (.03)	< .001
Outcome × Stake size	.03 (.02)	.17
Transition × Stake size	.00 (.02)	.97
Outcome × Transition × Stake size	-.00 (.02)	.85

References

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

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*(12), 1704-1711.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience, 18*(5), 767-772.

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology, 71*, 1-6.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66*, 585-595.

Rummery, G., & Niranjan, M. (1994). On-line Q-learning using connectionist systems. *Cambridge University*.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.