Supplemental Materials

Supplemental Methods

Computational Model. We fit behavior to a computational model described in prior work (Kool, Gershman, & Cushman, 2017). This model and its theoretical background have been described extensively in previous work; here, we briefly outline its instantiation in the present research.

The model assumes that participants employ a hybrid of model-based and model-free learning. At the second stage of a trial in the two-step task (i.e., revealing the stock payoff), model-based and model-free values are equivalent, given that there are no further state transitions (e.g., additional stimuli) following second-stage rewards. As a result, for each second-state *S2,* state-action values *Q2* are learned for each action *a2* (i.e., selecting the stock to reveal a payout). These values are updated on each trial *t* after rewards are received, according to:

$$Q2(s2_t, a2_t) = Q2(s2_t, a2_t) + \alpha\delta_{2,t}$$

where $\delta_{2,t}$ is the prediction error at the second stage:

$$\delta_{2,t} = r_t - Q2(s2_t, a2_t)$$

given reward *r* and learning rate α, which controls the impact of new rewards on values.

At the first stage, model-based and model-free values differ. The model-free system uses SARSA(λ) (state-action-reward-state-action temporal difference learning algorithm with eligibility trace parameter λ) to learn a value, *QMF*, for each action *a1* at each first stage *s1* (i.e., values for choosing each advisor). These values are updated after each stage. Following the first stage, no direct rewards are experienced, and so model-free values for advisors are updated based on the value of the second stage action:

$$QMF(s1_t, a1_t) = QMF(s1_t, a1_t) + \alpha\delta_{1,t}$$

where

$$\delta_{1,t} = Q2(s2_t, a2_t) - QMF(s1_t, a1_t)$$

Following the second stage action, model-free values for advisors are updated based on the rewards experienced:

$$QMF(s1_t, a1_t) = QMF(s1_t, a1_t) + \lambda\alpha\delta_{2,t}$$

where λ is an eligibility trace parameter that down-weights the impact of the second-stage prediction error on first-stage values.

The model-based system uses the task transition structure to choose an advisor at Stage 1, planning ahead based on the *Q2* values of each stock. Formally, this depends on the probability of transitioning to each second-stage state and the reward available for the best action in each second-stage state (see Kool et al., 2017; Doll et al., 2015). Since transitions in the current task were deterministic and there was only one action available at the second stage, this simplifies in the present task to:

$$Q_{MB}(s1_t, a1_t) = Q2(S(s1_t, a1_t), A(s1_t, a1_t))$$

where $S(s1_t, a1_t)$ is the second-stage state that would be produced by choosing action *a1* in the first-stage state *s1* and $A(s1_t, a1_t)$ is the action that would be available in that state. (Note that only one action was available in each second state; one stock was shown onscreen, and participants pressed a key to reveal the payout.)

To make a choice at the first stage, the model combines model-based and model-free values using a weighting parameter *w:*

$$Q_{net}(s1_t, a1_t) = wQ_{MB}(s1_t, a1_t) + (1 - w)Q_{MF}(s1_t, a1_t)$$

Note that if $w = 0$, choices would be fully model-free, and if $w = 1$, choices would be fully model-based.

Finally, choices were modeled using a softmax choice function:

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

where β is an inverse temperature parameter controlling the stochasticity of choice, *rep(a)* indicates if first-stage action *a* is the same one chosen on the previous trial (1 if yes, 0 if no), *resp(a)* indicates if the same response key was pressed as on the previous trial (1 if yes, 0 if no), and π and ρ are "stickiness" and "response stickiness" parameters that weight these indicators, respectively. The "stickiness" terms control the extent to which participants show perseveration in choosing the same stimulus (i.e., the same advisor) and perseveration in making the same button press (e.g., pressing the button on the right).

Model-fitting. Parameters were estimated using maximum a posteriori (MAP) estimation with empirical priors (Gershman, 2016). Following Kool et al (2017), we used priors of $\beta \sim$ Gamma(4.82, 0.88), and stickiness parameters, $\pi, \rho \sim \mathcal{N}(0.15, 1.42)$; other parameters had flat priors. The learning rate, weighting parameter, and eligibility trace decay were bounded between 0 and 1; the inverse temperature parameter was bounded between 0 and 20; and the stickiness parameters were bounded between -20 and 20.

Model comparison. Our parameter fits revealed a high degree of model-based learning (Table S1); indeed, 20 participants out of 65 had fitted *w* parameters equal to 1. As a complementary approach to test the impact of both model-free and model-based learning, we fit an alternate model in which the *w* parameter was constrained to 1 (fully model-based) for all subjects. We compared this model to the hybrid model—which contains an additional free parameter—using Bayesian model selection (Stephan, Penny, Daunizeau, & Moran, 2009) as implemented in the m-fit package for Matlab (Gershman, 2016). Surprisingly, this procedure supported the constrained model ($w = 1$), exceedance probability = 1.

Given that the regression analysis showed positive evidence of model-free learning—an interaction of Reward x Start State that would not be produced by model-based learning alone—it is possible that the model comparison failed to select the hybrid model given high *w* parameters on average. To test this possibility, we conducted a series of simulations. We simulated the task 100 times using the exact parameter fits from our subjects, with each simulated subject's parameters corresponding to one real subject's parameters, and we next simulated the task 100 times using the same parameter fits but with the *w* parameter fixed to 1 for each subject. For each simulation, we conducted the regression and model comparison analyses. This procedure allowed to ask how well each analysis detected effects under each ground truth (*w* varying vs $w = 1$ across the full sample) and allowed us to rule out the possibility that the regression analysis might produce an artifactual model-free effect even when $w = 1$ in all subjects.

When *w* was set to 1 for all subjects, the regression analysis returned a significant model-free effect in only 3% of simulations, close to an alpha rate of .05. Moreover, the model comparison procedure supported the constrained model ($w = 1$) in all iterations. This simulation indicates that the regression analysis would not artifactually produce evidence of model-free learning if the true *w* were equal to 1 in all subjects.

When *w* was high but less than 1 in most subjects (using the actual fits from our sample), the model-comparison procedure still favored the constrained model in all iterations. This result demonstrates that model comparison may fail to select the hybrid model if the *w* parameter is quite high on average. However, the regression analysis recovered a significant model-free effect in 53% of simulations, suggesting that this analysis can correctly recover a model-free effect, although it was somewhat underpowered to do so.

Altogether, these simulations support the conclusions in the main text: the regression analysis revealed the influence of model-free learning in choice on average, although it is possible that this effect was not present in each individual subject.

Generative model for plots. To generate the model predictions shown in Figure 2, we simulated the task 100 times with 65 subjects given the model described above. To do so without drawing on our own data, we used the median parameter fits reported by Kool et al. (2017), changing only the w parameter to 0 or 1 to simulate fully model-free or model-based behavior, respectively.

References

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.

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

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, *28*(9), 1321-1333.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, *46*(4), 1004-1017.

Supplementary Tables

Table S1.

Parameter fits for the computational model.

Statistic	β	α	λ	w	π	ρ
25th percentile	.58	.80	0	.70	-.08	-.40
Median	.70	1	.51	.92	.27	-.07
75th percentile	.90	1	.75	1	.86	.19

Table S2.

Coefficients for regression analysis of learning phase choice.

Effect	**Coefficient**	***SE***	***z***	***p***
Intercept	1.70	0.09	17.98	$< 2.00 \times 10^{-16}$
Reward	1.47	0.07	19.79	$< 2.00 \times 10^{-16}$
Same State	0.20	0.04	4.81	1.49×10^{-06}
Reward x Same State	0.22	0.03	6.45	1.14×10^{-10}

Table S3.

Coefficients for regression analysis of post-learning evaluations. The weighting parameter was z-scored across participants prior to analysis; main effects of model-free values (QMF) and model-based values (QMB) can therefore be interpreted at the mean level of the weighting parameter (w = .83).

Effect	**Coefficient**	***SE***	***df***	***t***	***p***
Intercept	4.73	0.11	63.39	44.64	$< 2.00 \times 10^{-16}$
QMF	0.16	0.09	162.97	1.82	0.07
QMB	0.30	0.14	71.46	2.17	0.03
W	0.07	0.11	64.45	0.63	0.53
QMF x W	-0.24	0.08	148.01	-2.97	0.004
QMB x W	-0.12	0.14	67.78	--0.90	0.37