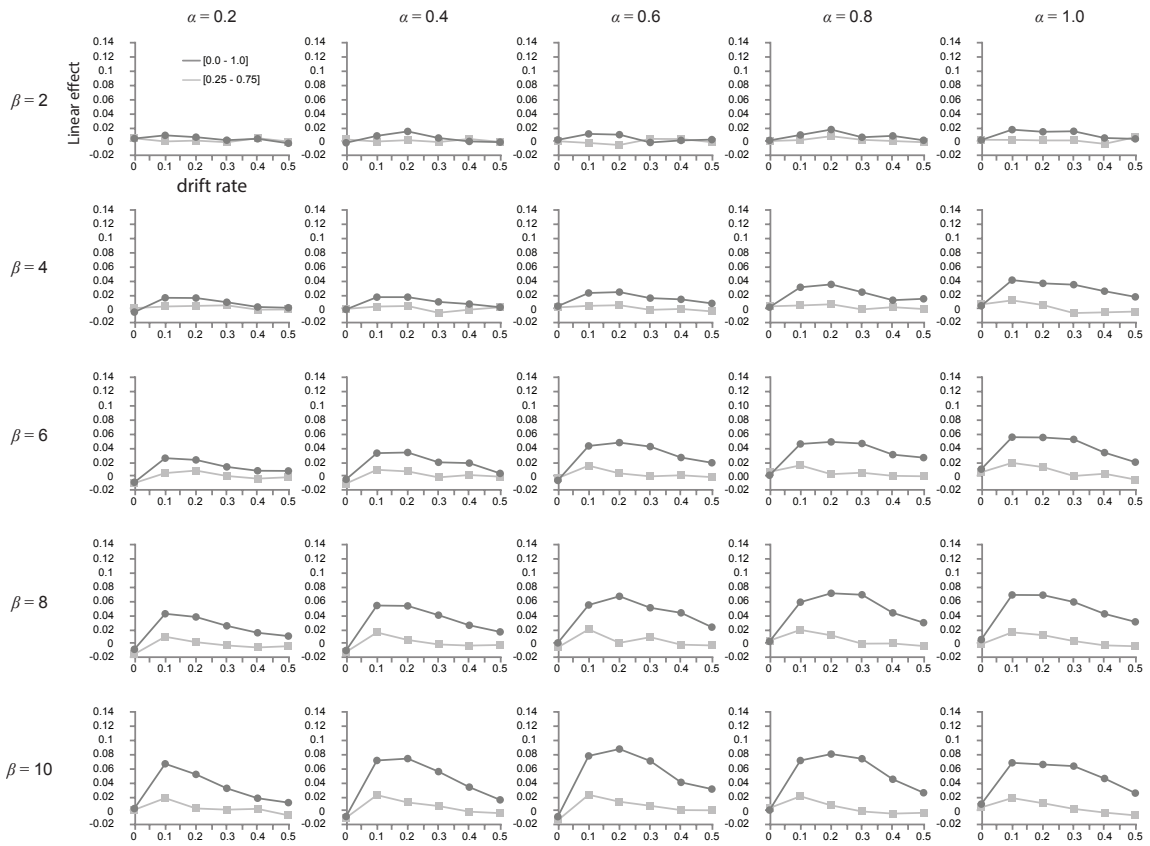
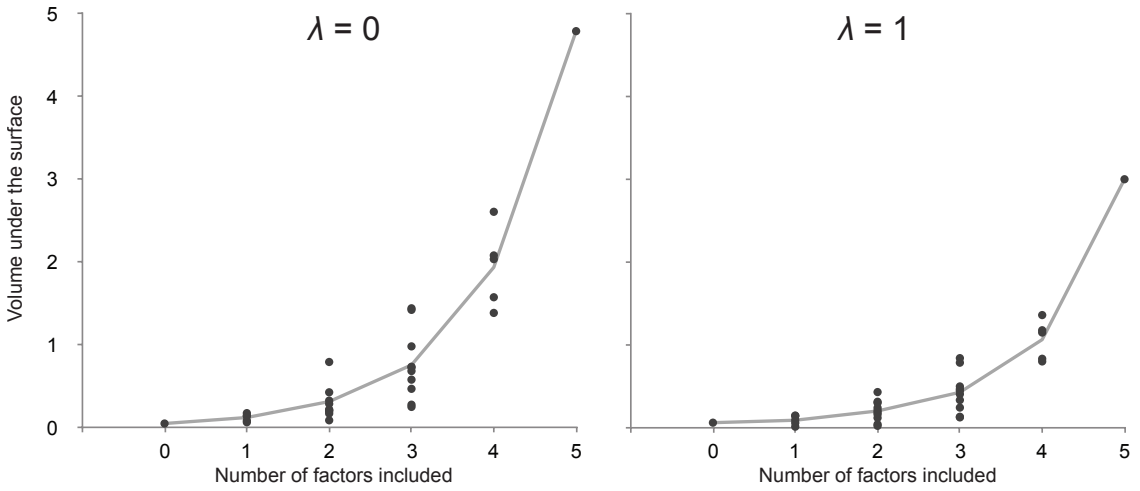


S1 Fig. The influence of the drift rate in the two-step task across a broad range of RL parameters. We found that the size of the drift rate affected the strength between model-based control and reward in a non-monotonic fashion, with the largest effect found at moderate values of the drift rate (0.1-0.3) and with a broad reward probability range. Importantly, the results of this analysis shows that this effect was not only found in the particular parameterization depicted in Figure 8 in the main text, but also across a broad range of learning rates (α) and inverse temperatures (β).



S2 Fig. Volume under the surface for all 32 tasks generated by the 5 binary factors discussed in this paper for agents with eligibility decay parameter λ = 0 and λ = 1. Each dot represents the volume under the surface of linear regression coefficients for one task, and is plotted as a function of the number of 'beneficial' factors that are included in each task's design. The gray line represents the average increase in the strength of the relationship between model-based control and reward. These results are qualitatively identical to those reported in Figure 13, suggesting that λ does not reliably affect the strength of the accuracy-efficiency tradeoff.

S1 Text. Reliability analysis for model-fitting procedure

Here, we report an analysis to test whether our model-fitting procedure could reliably dissociate model-based from model-free control (even though there was a lack of qualitative difference in single-trial staying behavior, see Figure 13B in the main text). In order to do so, we used the generative RL model of our novel two-step paradigm to simulate behavioral performance for 200 agents with randomly selected parameters. For each agent, we randomly sampled parameters from uniform distributions: $\{\alpha, \lambda, w\} \sim U(0, +1)$, $\beta \sim U(0, +2)$.

Next, we used our model-fitting procedure to estimate parameters for each of our agent using MATLAB's *patternsearch* function. To avoid local optima in the estimation solution, we ran 25 iterations for each agent with randomly selected starting positions for each parameter. We extracted from the fit with maximal log-likelihood. The final estimations for all parameters were extracted from the iteration with the maximal log-likelihood.

We found substantial correlations between the true and estimated values for submitted parameters. Most importantly, the weighting parameter w showed a strong and positive correlation, $r = 0.68$, $p < 0.001$, showing that our method is able to extract meaningful parameter estimates even when behavioral data do not qualitatively discriminate between different modes of processing. Correlations for the other parameters were as follows. Inverse temperature parameter β, $r = 0.25$, $p < 0.001$, learning rate parameter α, $r = 0.82$, $p < 0.001$, and eligibility trace decay parameter λ, $r = 0.27$, $p < 0.001$.

S1 Table. Model comparison for the full hybrid model and the hybrid model without choice perseveration parameters.

Paradigm	Model	Number of parameters (*k)*	Log-likelihood	BIC	AIC	McFadden's pseudo R^2
Daw	No stickiness	4	-27541	58886	56657	0.19
	π	5	-26082	56920	54134	0.24
	ρ	5	-26569	57893	55107	0.22
	π , ρ	**6**	-25086	**55880**	**52537**	0.27
Novel	**No stickiness**	**4**	-11742	27038	24957	0.26
	π	5	-11535	27511	24909	0.28
	ρ	5	-11273	**26989**	24387	0.29
	π , ρ	6	-11035	27401	**24279**	0.31

Note: The number of trials in both experiments was $n = 125$, and the number of participants in the Daw paradigm $N = 197$, and in the novel paradigm, $N = 184$, and therefore BIC = -2 × Log-likelihood + $k \times N \times \log(n)$, and AIC = $2 \times k \times N$ - 2 × Log-likelihood. McFadden's pseudo R^2 is computed as (R - Log-likelihood)/R where R is the log-likelihood for the chance model ($125 \times 2 \times \ln(1/2)$ for the Daw paradigm and $125 \times \ln(1/2)$ for the Doll paradigm).

S2 Table. Model comparison for the hybrid model and pure model-based and model-free models.

Paradigm	Model	Number of parameters (*k*)	Log-likelihood	BIC	AIC	McFadden's pseudo R^2
Daw	Hybrid	6	-25086	55880	**52537**	0.27
	Model-free	5	-25346	**55449**	52663	0.26
	Model-based	5	-25561	55878	53092	0.25
Novel	**Hybrid**	5	-11273	**26989**	**24387**	0.29
	Model-free	4	-12173	27900	25818	0.24
	Model-based	4	-11804	27162	25080	0.26

Note: The number of trials in both experiments was $n = 125$, and the number of participants in the Daw paradigm $N = 197$, and in the novel paradigm, $N = 184$, and therefore BIC = $-2 \times \ln(\text{Likelihood}) + k \times N \times \ln(n)$, and AIC = $2 \times k \times N - 2 \times \ln(\text{Likelihood})$. McFadden's pseudo R^2 is computed as $(R - \ln(\text{Likelihood}))/R$ where R is the log-likelihood for the chance model ($125 \times 2 \times \ln(1/2)$ for the Daw paradigm and $125 \times \ln(1/2)$ for the Doll paradigm).

S2 Text. Multilevel logistic regression analyses

In addition to the model fitting analyses reported in the main text, we also fit multilevel logistic regression models to the data from both participants that completed the Daw paradigm and those that completed the novel paradigm. These analyses were carried out using the lme4 package (http://cran.r-project. org/web/packages/lme4/index.html) in the R statistical language (http://www. r-project.org/). In order to measure individual difference in choice behavior, we modeled all coefficients as random effects, varying between participants around a group mean.

Analysis

Novel paradigm

For participants in the novel paradigm, we predicted whether the participants repeated the previous trial's second-stage state (i.e., "staying behavior") as a function of the rewards and similarity of the previous trial's first-stage state. Specifically, the dependent variable was whether the current second-stage choice was the same as that on the previous trial. For each trial i, the predictors for this analysis were the amount of points on the previous trial (r_{i-1}) and whether the previous starting state was the same or different from the current starting state ($same_i = 1$ or $same_i = 0$, respectively). In addition, in order to account for the influence of the prior reward history, we included a predictor that coded for the difference in possible reward between chosen and unchosen terminal states on the previous trial ($difference_{i-1}$). The final multilevel regression model included these three predictors, their interactions and the intercept. We ran this analysis for the experimental data ($n = 185$). In addition, in order to gain more insight into the range of possibilities for this analysis, we ran the same analysis for two simulated sets of data that were generated using the dual-systems RL model of the novel paradigm. For each set, we matched one

agent to each of our participants, copying their parameter fits except for the weighting parameter, which was set to $w = 0$ for one set (the model-free set), and $w = 1$ for the other (the model-based set).

In these regression analyses, the main effect of the previous reward (r_{i-1}) represents the model-based contribution to choice, since it carries over to the next trial even when the start states are different ($same_i = 0$), whereas the interaction term $r_{i-1} * same_i$ captures reward effects that are specific to the state in which they were received and therefore represent the model-free contribution to choice. We computed the difference between the model-based and model-free coefficients as an analogous term to the weighting parameter w in the computation model.

Daw paradigm

For the Daw task, we 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}) and what type of transition occurred on that trial ($common_i$). The final multilevel regression model included these two predictors, their interactions and the intercept. We ran this analysis for the experimental data ($n = 198$). We also again ran the same analysis for two simulated sets of data that were generated using the dual-systems RL model of the Daw paradigm. For each set, we matched one agent to each of our participants, copying their parameter fits except for the weighting parameter, which was set to $w = 0$ for one set (the model-free set), and $w = 1$ for the other (the model-based set).

In these regression analyses, the main effect of the previous reward (r_{i-1}) represents the model-free contribution to choice, since it captures reward effects that are independent from the transition type on the previous trial, whereas the whereas the interaction term $r_{i-1} * common_i$

captures reward effects that are modulated by the transition type on the previous trial and therefore represents the model-based contribution to choice. For the experimental data, we computed the difference between the model-based and model-free coefficients as an analogous term to the weighting parameter *w* in the computational model.

Results

Novel paradigm. The results from the logistic regression for the novel paradigm are given in Table S3. For the experimental data we found significant effects of the regressors indicating the outcome of the previous trial, indicating a model-based contribution, and the interaction between previous outcome and the similarity of the current and previous first-stage states, indicating the model-free contribution. Importantly, we found no significant interaction between reward and similarity of the current and previous first-stage states for the simulated data of model-based agents, but this interaction was significant for the model-free agents, as expected.

Table S3. Regression coefficients from multilevel logistic regression analysis, indicating the effect of outcome of previous trial, similarity of previous starting state to current starting state, and previous difference between chosen and unchosen reward, on repetition of second-stage choice for experimental data and simulated data from pure model-based and pure-model free agents matched to fits from experimental data.

Coefficient	Experimental data		Model-free agents		Model-based agents	
	Estimate (SE)	*p*	Estimate (SE)	*p*	Estimate (SE)	*p*
(Intercept)	.47 (.04)	< .001	.42 (.04)	< .001	.54 (.05)	< .001
Previous reward	.31 (.02)	< .001	.13 (.01)	< .001	.23 (.02)	< .001
Same starting state	.21 (.03)	< .001	.15 (.02)	< .001	-.00 (.02)	.98
Previous reward difference	.06 (.01)	< .001	.04 (.01)	< .001	.08 (.01)	< .001
Reward × Same	.16 (.02)	< .001	.07 (.01)	< .001	-.00 (.01)	.71
Reward × Difference	-.00 (.00)	.39	-.00 (.00)	< .01	.01 (.00)	< .01
Difference × Same	-.00 (.01)	.98	.00 (.01)	.79	.00 (.00)	.97
Reward × Same × Difference	.01 (.00)	.02	-.00 (.00)	.76	.00 (.00)	.87

Daw paradigm. The results from the logistic regression for the Daw paradigm are given in Table S4. For the experimental data we found significant effects of the regressors indicating the outcome of the previous trial, indicating a model-free contribution, and the interaction between previous outcome and previous transition type, indicating the model-free contribution. For the simulated data, we found no significant interaction between reward and transition type for the simulated data of model-free agents and no significant main effect of reward for the simulated data of model-based agents as expected.

Table S4. Regression coefficients from multilevel logistic regression analysis, indicating the effect of outcome of previous trial, transition type of previous trial, on repetition of first-stage choice for experimental data and simulated data from pure model-based and pure-model free agents matched to fits from experimental data.

	Experimental data		Model-free agents		Model-based agents	
Coefficient	Estimate (SE)	*p*	Estimate (SE)	*p*	Estimate (SE)	*p*
(Intercept)	1.03 (.07)	< .001	1.19 (.09)	< .001	.80 (.06)	< .001
Previous reward	.26 (.03)	< .001	.33 (.03)	< .001	.01 (.02)	.54
Previous transition	.03 (.02)	.14	.00 (.02)	.89	.03 (.02)	.12
Reward × Transition	.20 (.03)	< .001	.02 (.02)	.38	.18 (.02)	< .001

Correlations. We found that our indices of the relative weighting between model-based and model-free control were positively related to our measure of reward that controlled for average chance performance for the novel task ($r = 0.69$, $p < 0.001$), but not for the Daw paradigm ($r = 0.03$, $p = 0.71$). A subsequent multiple regression showed that this relationship was significantly different between groups [$t(377) = 7.33$, $p < 0.001$]. These results provide convergent evidence for the accuracy-demand trade-off of the novel two-step paradigm, and for the verification of the prediction that the original Daw two-step paradigm does not embody such a trade-off.