Supporting Information

Otto et al. 10.1073/pnas.1312011110

SI Text

Reinforcement Learning Model. The task consists of three states (first stage: s_A; second stage: s_B and s_C), each with two actions (a_A and a_B). The hybrid model consists of model-based and model-free subcomponents, both of which estimate a state-action value function $Q_{MF}(s,a)$ (model-free) and $Q_{MB}(s,a)$ (model-based) that maps each state-action pair to its expected future reward. On trial t, we denote the first-stage state (always s_A) by $s_{1,t}$, the second-stage state by $s_{2,t}$, the chosen first- and second-stage actions by $a_{1,t}$ and $a_{2,t}$, and the first- and second-stage rewards as $r_{1,t}$ (always zero) and $r_{2,t}$.

Model-free component. For the model-free algorithm, we used State-Action-Reward-State-Action, SARSA(λ), temporal difference (TD) learning (1), which updates the value for the visited state-action pair at each stage i and trial t according to

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

where

$$\delta_{i,t} = \left[r_{i,t} + Q_{MF}(s_{i+1,t},a_{i+1,t})\right]/\alpha - Q_{MF}(s_{i,t},a_{i,t}),$$

is the reward prediction error (RPE), and α is a learning rate parameter. For the first-stage choice, $r_{1,t}=0$ and the RPE is driven by the second-stage value, $Q_{MF}(s_{2,t},a_{2,t})$; conversely, at the second stage, we define $Q_{MF}(s_{3,t},a_{3,t})=0$, because there is no further value in the trial apart from the immediate reward $r_{2,t}$. Here we have rescaled the leading term in the reward prediction error by $1/\alpha$, relative to its usual definition (2, 3). Because this simply rescales the units of the Q values (by $1/\alpha^2$ and $1/\alpha$ at the first and second stage, respectively), the same data likelihoods are maintained via a corresponding rescaling of the first- and second-level inverse temperatures β_{MF} and β_2 in the choice rule below. This slight reparameterization facilitates group-level modeling by reducing the correlation of the β s with α.

The model uses an eligibility trace to propagate second-stage reward information to the first-stage values. Specifically, at the end of each trial, the first-stage values are updated according to

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

where λ is an eligibility trace decay parameter (4), and the omission of α (which would normally appear in this equation) again results from rescaling the update to match the scaling implied by the prediction error above. We assume that eligibility traces are reset to 0 between episodes (i.e., that eligibility does not carry over from trial to trial).

Additionally, at the end of each trial, we decayed the Q values for all of the nonchosen actions by multiplying them by $1-\alpha$ (5, 6). This decay makes the present model correspond more closely to the one-trial-back regression model described in the main text, in the limit as $\alpha \to 1$.

Model-based component. In general, a model-based reinforcement learning (RL) algorithm works by learning a transition function (mapping state-action pairs to a probability distribution over the subsequent state), and immediate reward values for each state, then computing cumulative state-action values by iterative expectation over these. Specialized to the structure of the current task, this amounts to, first, simply deciding which first-stage action maps to which second-stage state (because subjects were instructed that this was the structure of the transition contingencies), and second, learning immediate reward values for each of the second-stage actions (the immediate rewards at the first stage being always zero).

Following ref. 7, we modeled transition learning by assuming subjects simply chose between the two possibilities: $P(s_B|s_A,a_A)=0.7$, $P(s_C|s_A,a_B)=0.7$, or vice versa, $P(s_B|s_A,a_A)=0.3$, $P(s_C|s_A,a_B)=0.3$, with $P(s_B|s_A,a_B)=1-P(s_B|s_A,a_A)$ and $P(s_C|s_A,a_A)=1-P(s_C|s_A,a_B)$, according to whether more transitions had thus far occurred to s_B following a_A plus s_C following a_B, or vice versa to s_C following a_A plus s_B following a_B.

At the second stage (the only one where immediate rewards were offered), the problem of learning immediate rewards is equivalent to that for TD above, because $Q_{TD}(s_{2,t},a_{2,t})$ is just an estimate of the immediate reward $r_{2,t}$; with no further stages to anticipate, and the SARSA learning rule reduces to a delta rule for predicting the immediate reward. Thus, the two approaches coincide at the second stage, and we define $Q_{MB}=Q_{TD}$ at those states.

Finally the top level model-based values are defined from the transition and reward estimates using the Bellman Equation (8):

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

where we have assumed these are recomputed at each trial from the current estimates of the transition probabilities and rewards.

Choice rule. Finally, to connect the values to choices, we use a softmax choice rule, which assigns a probability to each action according to the combination of both Q_{MB} and $Q_{MF,}$ each weighted with a separate inverse temperature parameter, and β_{MB} and β_{MF}, which allow the two values to combine independently in determining first-stage choice. (Note that this is algebraically equivalent to the formulation used in ref. 7, under the substitution $\beta_{MB}=w\beta$ and $\beta_{MF}=(1-w)\beta$. This change of variables again facilitates group level modeling of individual differences in the influence of either system.

The probability of a choice at the first stage is calculated, accordingly, as

$$P(a_{i,t}=a|s_{1,t}) = \frac{\exp\left[\beta_{MB}\cdot Q_{MB}(s_{1,t},a) + \beta_{MF}\cdot Q_{MF}(s_{1,t},a) + p\cdot rep(a)\right]}{\sum_{a'}\exp\left[\beta_{MB}\cdot Q_{MB}(s_{1,t},a') + \beta_{MF}\cdot Q_{MF}(s_{1,t},a') + p\cdot rep(a')\right]}.$$

The indicator function $rep(a)$ is defined as 1 if a is a top-stage action and is the same one as was chosen on the previous trial, zero otherwise. Together with the "stickiness" parameter p, this

captures first-order perseveration ($p > 0$) or switching ($p < 0$) in the first-stage choices. Second-stage choices are modeled with an analogous but simpler softmax rule, with only a single value term $Q_{MF}(s_{2,t}, a)$, with its own inverse temperature β_2 and omitting the $rep(a)$ term.

Group-level modeling. The foregoing describes the modeling of a single subject's data. This model was embedded within a multilevel random effects model to estimate it for all subjects simultaneously. All of the free parameters of the model (α, λ, β_{MB}, β_{MF}, β_2, p) were taken as random effects, instantiated separately for each subject s from a common group level distribution. For parameters with infinite support, the group level distributions were Gaussian with free mean and SD

$$\beta_{2_s} \sim N(\mu_{\beta 2}, \sigma_{\beta 2}),$$

and similarly for p_s. To test the dependence of the model-based and model-free effects on cortisol and Operation Span (OSPAN), these effects and their interaction were entered into a regression at the group level

$$\beta_{MB_s} \sim N\Big[\mu_{\beta MB} + \beta_{MBcort} cort(s) + \beta_{MBospan} ospan(s) + \beta_{MBcxo} cort(s) \cdot ospan(s)\Big],$$

and similarly for β_{MF_s}. Accordingly, nonzero values of the slopes β_{MBcort}, $\beta_{MBospan}$, and β_{MBcxo} signify correlations between cortisol delta, OSPAN, and the interaction between the two, analogous to the covariate effects tested in the logistic regression in the main text.

The parameters with support in $[0, 1]$ were assumed to be drawn from a group-level beta distribution

$$\alpha_s \sim Beta(A_\alpha, B_\alpha)$$

and similarly for λ_s.

Finally, we estimated the parameters of the group level distributions ($\mu_{\beta 2}$, etc.) using uninformative priors: for all means, the broad Gaussian $N(0, 100)$, for all SDs, the heavy-tailed $Cauchy(0, 2.5)$. Finally, our priors for the A and B parameters of the beta distributions were given using a change of variables that characterizes the distribution's mean $M = \frac{A}{A+B}$ and spread $S = \frac{1}{\sqrt{a+b}}$, the latter approximating its SD. This allowed us to take as uninformative hyperpriors the uniform distributions $M \sim U(0, 1)$ and $S \sim U(0, \infty)$ (the latter improper) (9).

Estimation. We estimated the joint distribution of the parameters of the model, conditional on all subjects' observed choices and rewards. For this, we used Markov Chain Monte Carlo (MCMC) techniques (specifically the No-U-Turn variant of Hamiltonian Monte Carlo) as implemented in the Stan modeling language (10). Given a probabilistic generative model (the above equations) and a subset of observed variables, MCMC techniques provide samples from the conditional joint distribution over the remaining random variables. We ran four chains of 2,000 samples each, discarding the first 1,000 samples of each chain for burn-in. We examined the chains visually for convergence and also computed Gelman and Rubin's (11) potential scale reduction factors. For this, large values indicate convergence problems, whereas values near 1 are consistent with convergence. We ensured that these diagnostics were less than 1.1 for all variables.

Results. Table S2 reports the free parameters of the model by their group-level means and variances over individual subjects. Also reported are the regression slopes estimating how individuals' parameter settings covaried with cortisol deltas, OSPAN scores, or their interaction. This uncertainty is reported via quartiles: the median and 25th and 75th percentiles of the distribution. Of note, the group-level mean α was centered on 0.34, characteristic of a more gradual (and thus, less recency driven) learning process than is ascribed by the regression analysis in the main text, which assumes a learning rate of 1 (that is, only the most recent trial influences choice), supporting the conclusion that our reported effects apply to longer-term incremental learning, and are not limited to short-term patterns of win-stay-lose-shift adjustments.

Regression Analysis. We specified a mixed-effects logistic regression to explain the first-stage choice on each trial t (coded as stay vs. switch) using binary predictors indicating if reward was received on t-1 and the transition type (common or rare) that had produced it. Logistic regressions were conducted as mixed-effects models, performed using the lme4 package (12) in the R programming language. Within-subject factors (the intercept, main effects of reward and transition, and their interaction) were taken as random effects across subjects, and estimates and statistics reported are at the population level. Individual model-based and model-free effect sizes (the model-based and model-free indices used in Figs. S1 and S2) were calculated from posterior estimates, conditional on the estimated top-level effects. Planned contrasts were conducted using the esticon function (package doBy) (13) on the estimated model.

As an initial examination, we estimated a model that included both experimental condition (stress v.s control) and cortisol delta as between subjects-factors (Table S3). Statistically, we found a significant negative interaction between cortisol response (quantified by cortisol delta; *Materials and Methods*), previous reward, and transition type ($P < 0.01$), confirming that cortisol response effectively attenuated the model-based signature of choice. Experimental condition (stress vs. control), however, did not exert significant influence on choice-related variables, nor did it significantly interact with the interaction between cortisol response and these trial-by-trial variables. That cortisol response yields greater explanatory leverage on behavior than experimental condition mirrors the results of recent examinations of stress and decision-making (14, 15). A separate regression, excluding condition, is reported in Table 2. Further, this regression confirmed that cortisol response did not influence the simple effect of previous reward—the hallmark of model-free learning ($P > 0.5$). Moreover, the effect of cortisol response on model-based contributions trended larger than model-free contributions (linear contrast between the reward effect and the reward × transition interaction, $P = 0.07$), positively demonstrating the selectivity of the effect to model-based RL and suggesting that cortisol response does not merely bring about a generalized decline in performance.

To visualize the relationship between cortisol response and model-based contribution to behavior analogously to the computational model weights, we computed for each subject a model-based index (the individual's coefficient estimate for the previous reward × transition type interaction as in Fig. 5*A*, the marker of model-based updating). Fig. S1*A* plots the model-based index as a function of cortisol response and condition, suggesting that the model-based contribution to choice decreased as a function of cortisol increase. Plotting the model-free index, an individual measure of the model-free contribution to choice (the coefficient for the main effect of the previous trial's reward on choice), as a function of cortisol response and condition revealed no apparent attenuation of model-free choice by cortisol response or condition (Fig. S1*B*).

We applied the same analysis approach to examine how individual working-memory (WM) capacity—operationalized by OSPAN—modulates the effect of cortisol response on model-based choice. Accordingly, we examined this relationship with a logistic model examining how cortisol delta and OSPAN interacted with the same trial-by-trial variables in the above analysis (previous reward and transition type; see Table S4 for full model specification and coefficient estimates). Critically, OSPAN significantly interacted with the three-way interaction between cortisol re-

sponse, previous reward, and previous transition type (the interaction signifying cortisol response's effect on model-based choice, $P < 0.01$). This relationship is visualized in Fig. S2, analogous to Fig. 4: among subjects low in WM capacity, cortisol delta reduced the expression of model-based choice (Fig. 4*A*), but among subjects high in WM capacity, cortisol response did not produce an appreciable impact on model-based contributions to behavior (Fig. 4*B*).

1. Rummery GA, Niranjan M (1994) *On-Line Q-Learning Using Connectionist Systems* (Cambridge Univ, Cambridge, UK).
2. Camerer C, Ho T-H (1999) Experienced-weighted attraction learning in normal form games. *Econometrica* 67(4):827–874.
3. Den Ouden HEM, et al. (2013) Dissociable Effects of Dopamine and Serotonin on Reversal Learning. *Neuron* 80:1090–1100.
4. Sutton RS, Barto AG (1998) *Reinforcement Learning* (MIT Press, Cambridge, MA).
5. Lau B, Glimcher PW (2005) Dynamic response-by-response models of matching behavior in rhesus monkeys. *J Exp Anal Behav* 84(3):555–579.
6. Ito M, Doya K (2009) Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *J Neurosci* 29(31):9861–9874.
7. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69(6):1204–1215.
8. Bellman R (1957) *Dynamic Programming* (Princeton Univ Press, Princeton, NJ).
9. Gelman A, Carlin JB, Stern HS (1995) *Bayesian data analysis* (Chapman & Hall, London).
10. Stan Development Team (2013) *Stan: A C++ Library for Probability and Sampling, Version 1.3.* Available at http://mc-stan.org/. Accessed November 20, 2013.
11. Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Stat Sci* 7(4):457–472.
12. Pinheiro JC, Bates DM (2000) *Mixed-Effects Models in S and S-PLUS* (Springer, New York).
13. Højsgaard S, Halekoh U (2009) *doBy: Groupwise Computations of Summary Statistics, General Linear Contrasts and Other Utilities.* Available at http://CRAN.R-project.org/package=doBy. Accessed November 20, 2013.
14. Starcke K, Polzer C, Wolf OT, Brand M (2011) Does stress alter everyday moral decision-making? *Psychoneuroendocrinology* 36(2):210–219.
15. Leder J, Häusser JA, Mojzisch A (2013) Stress and strategic decision-making in the beauty contest game. *Psychoneuroendocrinology* 38(9):1503–1511.

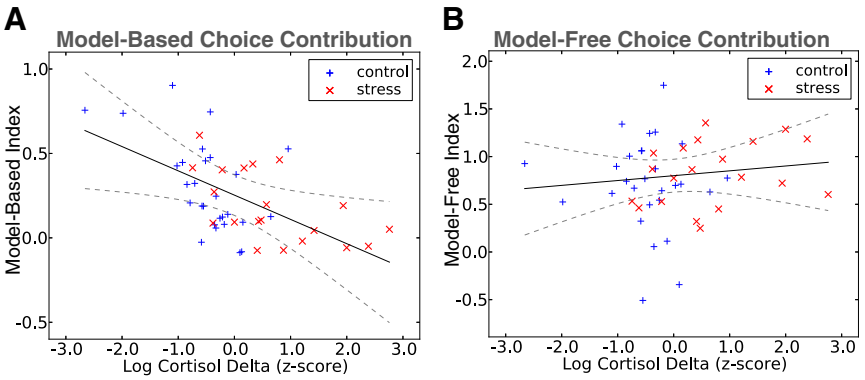

Fig. S1. Effect of cortisol response on model-based vs. model-free behavioral contributions. (*A*) Individual subjects' model-based effect sizes (arbitrary units) conditional on the group-level mixed-effects logistic regression, plotted separately for subjects in the control and stress conditions. The regression line is computed from the group-level log-linear effect of cortisol delta. There was a significant negative effect of cortisol delta on expression of model-based choice ($P < 0.05$), indicating cortisol change diminished its behavioral expression. (*B*) Subject-level effect-sizes for the model-free contribution to behavior. Note that there was no significant effect of cortisol change on expression of model-free choice ($P = 0.54$), indicating that expression of model-free contribution is spared. Dashed gray lines indicate 2 SEs, estimated from the group-level mixed effects regression.

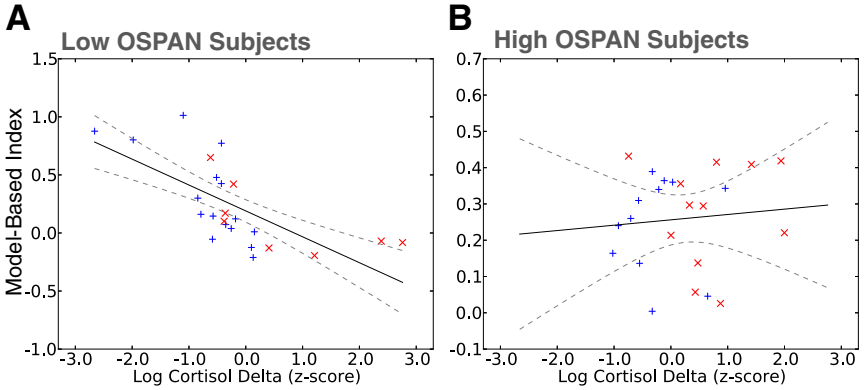

Fig. S2. Effect of cortisol response on expression of model-based behavior as a function of individual WM capacity as measured by OSPAN. Individual subjects' model-based effect sizes (arbitrary units) are plotted for low OSPAN subjects (*A*) and high OSPAN subjects (*B*). Cortisol response markedly dampened expression of model-based choice in the low OSPAN subgroup but not in the high OSPAN subgroup. Regression lines are computed from the group-level log-linear effect of cortisol delta for each subgroup. Dashed lines indicate 2 SEs, estimated from the group-level mixed effects regression.

Table S1. Mean cortisol response by group and sample time

	Sample			
Condition	t1 (baseline)	t2 (post-OSPAN)	t3 (post-CPT)	t4 (post-RL task)
Control (n = 28)	6.03 (3.08)	5.42 (2.47)	4.79 (2.31)	4.64 (2.65)
Stress (n = 20)	5.09 (3.54)	5.20 (3.01)	9.06 (6.49)	11.20 (13.26)

Salivary concentrations reported in nmol/L and are non–log transformed for interpretability.

Table S2. Group level estimates for the free parameters of the RL model and estimated slopes for the covariates

	Group-level means					
Percentile	β_{MB}	β_{MF}	p	β_2	λ	α
25	0.252	0.642	1.294	1.345	0.964	0.32
50	0.313	0.693	1.406	1.404	0.978	0.341
75	0.37	0.757	1.511	1.475	0.989	0.362
	Group-level variances					
	β_{MB}	β_{MF}	p	β_2	λ	α
25	0.389	0.458	0.922	0.590	0.016	0.193
50	0.436	0.504	0.996	0.643	0.038	0.206
75	0.496	0.554	1.077	0.711	0.069	0.221
	Covariate slopes					
	β_{MBcort}	$\beta_{MBospan}$	β_{MBcxo}	β_{MFcort}	$\beta_{MFospan}$	β_{MFcxo}
25	−0.285	0.051	0.318	−0.05	−0.061	−0.058
50	−0.226	0.113	0.42	0.007	−0.001	0.021
75	−0.163	0.17	0.525	0.063	0.062	0.108

For each parameter, the median posterior estimate is given, together with the quartiles of the posterior distribution. Note that the quartiles represent the width of uncertainty about the parameters' values (analogous to SEM), whereas the variances are estimates of the variability in the parameter estimates across the group of subjects.

Table S3. Logistic regression coefficients indicating the influence of cortisol response, stress condition, outcome of previous trial, and transition type of previous trial, on response repetition

Coefficient	Estimate (SE)	P value
(Intercept)	1.76 (0.20)	<0.0001*
Reward	0.72 (0.10)	<0.0001*
Transition	0.08 (0.07)	0.291
Cortisol delta	0.17 (0.33)	0.927
Condition	−0.11 (0.18)	0.690
Reward × transition	0.28 (0.07)	0.002*
Cortisol delta × reward	0.10 (0.18)	0.946
Cortisol delta × transition	0.02 (0.13)	0.170
Condition × reward	−0.09 (0.10)	0.702
Transition × cortisol delta	0.03 (0.07)	0.867
Condition × cortisol delta	0.20 (0.33)	0.391
Cortisol delta × reward × transition	−0.37 (0.13)	0.006*
Condition × reward × transition	0.11 (0.07)	0.321
Condition × cortisol delta × reward	0.18 (0.18)	0.314
Condition × cortisol delta × transition	0.06 (0.13)	0.347
Condition × cortisol delta × reward × transition	0.11 (0.13)	0.245

Critically, the cortisol delta × reward × transition was significant in the negative direction, indicating that cortisol response tempered model-based contribution to choice.
*Significance at the 0.05 level.

Table S4. Logistic regression coefficients indicating the influence of Operation Span (OSPAN) cortisol response, outcome of previous trial, and transition type of previous trial, on response repetition

Coefficient	Estimate (SE)	*P* value
(Intercept)	1.87 (0.17)	<0.0001*
Reward	0.77 (0.09)	<0.0001*
Transition	0.01 (0.05)	0.885
Cortisol delta	0.00 (0.16)	0.994
OSPAN	0.21 (0.17)	0.226
Reward × transition	0.20 (0.06)	<0.0001*
Cortisol delta × reward	0.03 (0.09)	0.734
Cortisol delta × transition	−0.08 (0.05)	0.090
OSPAN × reward	0.08 (0.09)	0.392
Transition × cortisol delta	0.08 (0.04)	0.084
OSPAN × cortisol delta	−0.12 (0.24)	0.633
Cortisol delta × reward × transition	−0.17 (0.06)	0.004*
OSPAN × reward × transition	0.09 (0.06)	0.099
OSPAN × cortisol delta × reward	0.06 (0.13)	0.619
OSPAN × cortisol delta × transition	0.35 (0.07)	<0.0001*
OSPAN × cortisol delta × reward × transition	0.23 (0.09)	0.009*

*Significance at the 0.05 level.