Supplemental Information

Pay-off structure of the gambling task

The number of points paid off by slot machine *i* on trial *t* ranged from 1 to 100, drawn from a Gaussian distribution (standard deviation σ_o = 4) around a mean $\mu_{i,t}$ and rounded to the nearest integer. On each trial, the means diffused in a decaying Gaussian random walk:

$$\mu_{i,t+1} = \lambda\mu_{i,t} + (1-\lambda)\theta + \nu$$

The decay parameter λ was 0.9836, the decay center θ was 50, and the diffusion noise ν was zero-mean Gaussian (standard deviation σ_d = 2.8). We used one instantiation of this process (Figure 1).

Reinforcement-learning model

We used a Bayesian mean-tracking rule (i.e., a Kalman filter) that tracked the mean expected pay-off of each machine ($\hat{\mu}_{i,t}$) and the variance of these pay-offs ($\hat{\sigma}^2_{i,t}$). On the first trial of the task, all four machines had the same prior mean $\hat{\mu}^{pre}_{i,1}$ and variance $\hat{\sigma}^{2\,pre}_{i,1}$. These start values were based on the pay-offs received during the practice block, and were determined separately for each participant (mean $\hat{\mu}^{pre}_{i,1}$ = 51.9, SD = 2.7; mean $\hat{\sigma}^{2\,pre}_{i,1}$ = 52.3, SD = 14.9). When a participant chose machine *c* on trial *t* and received pay-off *r*, the estimated pay-off distribution ($\hat{\mu}^{post}_{c,t}, \hat{\sigma}^{2\,post}_{c,t}$) was updated according to:

$$\hat{\mu}^{post}_{c,t} = \hat{\mu}^{pre}_{c,t} + \kappa_t\delta_t$$

$$\hat{\sigma}^{2\,post}_{c,t} = (1-\kappa_t)\hat{\sigma}^{2\,pre}_{c,t}$$

with prediction error $\delta_t = r_t - \hat{\mu}^{pre}_{c,t}$ and learning rate $\kappa_t = \hat{\sigma}^{2\,pre}_{c,t} / (\hat{\sigma}^{2\,pre}_{c,t} + \hat{\sigma}^2_o)$.

The estimated pay-off distributions for the unchosen machines did not change.

Then, the estimated prior pay-off distributions on the subsequent trial (trial *t*+1) were updated in time according to:

$$\hat{\mu}_{i,t+1}^{pre} = \hat{\lambda}\hat{\mu}_{i,t}^{post} + (1-\hat{\lambda})\hat{\theta}$$

$$\hat{\sigma}_{i,t+1}^{2\,pre} = \hat{\lambda}^2\hat{\sigma}_{i,t}^{2\,post} + \hat{\sigma}_d^2 .$$

We modeled the choice of the participants by a softmax rule. The probability $P_{i,t}$ of choosing machine *i* on trial *t* was given by:

$$P_{i,t} = \frac{\exp(\beta\hat{\mu}_{i,t}^{pre})}{\sum_j \exp(\beta\hat{\mu}_{j,t}^{pre})}$$

with exploration parameter β (often referred to as gain, or inverse temperature).

For a discussion of the Kalman filter and the softmax rule, we refer the reader to Anderson and Moore (1979), and Sutton and Barto (1998), respectively.

We fitted the model to each individual participant's choice data. The trials in which no response was made within the 1.5-s time limit were omitted. The parameters $\hat{\lambda}$, $\hat{\theta}$ and β were estimated per participant by maximizing the log-likelihood of the observed choices (Supplemental Table 2). Parameter $\hat{\sigma}_o$ was fixed at 4. Estimation of parameter $\hat{\sigma}_d$ resulted in extreme values for most of the participants (values larger than 1000 for ten of the seventeen participants), suggesting unreliable fits. Therefore, we fixed this parameter at 50, which is similar to the best fitting $\hat{\sigma}_d$ parameter found in a previous study (Daw, O'Doherty, Dayan, Seymour, and Dolan, 2006). This large value of $\hat{\sigma}_d$ implies that participants overestimate the speed of diffusion in the pay-offs. Large values of $\hat{\sigma}_d$ induce high learning rates, indicating that the expected pay-offs are determined primarily by the most recent experience with each machine.

Additional control analysis

Besides the multiple regression analyses, we performed a second set of control analyses to investigate whether differences in each of the potential confound variables could account for the different baseline pupil diameter on exploration and exploitation trials (and hence might provide an alternative interpretation of the effect). We repeated the comparison of baseline pupil diameter on exploitation and exploration trials while, in separate analyses, controlling for differences in each of the potential confound variables (pay-off on the previous trial, prediction error on the previous trial, expected pay-off on the current trial and entropy on the current trial), by matching the values of these variables across exploration and exploitation trials (Bernstein, Scheffers, & Coles, 1995). We sorted each participant's exploitation and exploration trials by one of these variables, and then successively removed the most extreme exploitation and exploration trials, thereby reducing the difference between the mean value of this confound variable on exploitation and exploration trials. After each trial removal, we calculated the difference between the mean values of the confound variable on exploitation and exploration trials, and we stopped the removal process when this difference was not further decreased by removal of a subsequent trial (Supplemental Table 1). We also controlled for choice strategy on the previous trial, by including only the trials that were preceded by an exploitation trial. Finally, in order to control more explicitly for the possibility that the higher incidence of negative prediction errors preceding exploratory choices was driving the effect, we repeated the analysis while only including the trials that were preceded by a positive prediction error. [†]

Importantly, none of the potential confound variables could account for the larger baseline pupils preceding exploratory compared to exploitative choices: the critical effect remained significant after correction for choice strategy on the previous trial [$t(16) = 2.5$, $p = 0.026$]; pay-off on the previous trial [$t(16) = 2.9$, $p = 0.009$]; prediction error on the previous trial [$t(16) = 2.5$, $p = 0.025$]; expected pay-off [$t(12) = 3.1$, $p = 0.010$]; and entropy [$t(16) = 3.5$, $p = 0.003$]. Furthermore,

the effect remained significant when only the trials that were preceded by a positive prediction error were considered ($t(12) = 2.3$, $p = 0.037$), suggesting that the larger baseline pupil on explore compared to exploit trials was not due to the larger incidence of negative prediction errors preceding explore trials.

Uncertainty-driven exploration and pupil diameter

In the softmax rule described above, the probability that a particular machine is chosen is determined by its relative mean estimated pay-off (and the value of the gain parameter), but not by the uncertainty about its potential pay-offs (i.e., the variance of the estimated pay-off distribution $\hat{\sigma}_i^{2\,pre}$). On the other hand, modeling studies have suggested that exploration might be directed towards particular choices in proportion to the uncertainty about their outcomes, which can be implemented by adding an 'uncertainty bonus' to the expected value of options with uncertain outcomes (e.g., Sutton, 1990). It has recently been shown that individual differences in uncertainty-based exploration are associated with the val158met polymorphism of the COMT gene, which substantially affects prefrontal dopamine levels (Frank, Doll, Oas-Terpstra, & Moreno, 2009). According to the adaptive gain theory, the increased NE level in the tonic LC mode indiscriminately facilitates processing of all stimuli and/or behaviors, which promotes a nonspecific type of exploration. Hence, the theory predicts that individual differences in tonic LC activity (as indexed by baseline pupil diameter in this study) will be related to individual differences in exploratory behavior (Results section), but not to individual differences in uncertainty-specific exploration.

To asses this last prediction, we considered a softmax rule in which an 'uncertainty bonus' of φ standard deviations was added to the estimated mean pay-offs:

$$P_{i,t} = \frac{\exp(\beta[\hat{\mu}_{i,t}^{pre} + \varphi\hat{\sigma}_{i,t}^{pre}])}{\sum_j \exp(\beta[\hat{\mu}_{j,t}^{pre} + \varphi\hat{\sigma}_{j,t}^{pre}])}$$

The best fitting uncertainty bonus parameter in this model varied across participants: four participants had a positive bonus and thirteen participants had a negative bonus (mean bonus = -0.117, SD = 0.336). Thus, for the majority of the participants, uncertainty about the potential outcomes of a machine *discouraged* exploration of that machine. Importantly, the value of the uncertainty bonus parameter did not correlate with baseline pupil diameter ($r = 0.05$, $p = 0.86$), consistent with the assumption that the tonic LC mode is not associated with uncertainty-specific exploration.

Footnote

† Four participants had to be excluded from the analysis that corrected for expected pay-off, because the difference in expected pay-off between their exploration and exploitation trials was so large that no exploration trials were left using this procedure. Similarly, four participants were excluded from the analysis in which only the trials preceded by a positive prediction error were considered, since less than ten explore and/or exploit trials were left for these participants.

Supplemental References

Anderson, B.D.O., & Moore, J.B. (1979). *Optimal Filtering.* Englewood Cliffs, NJ: Prentice-Hall.

Bernstein, P.S., Scheffers, M.K., & Coles, M.G. (1995). "Where did I go wrong?" A psychophysiological analysis of error detection. *Journal of Experimental Psychology. Human Perception and Performance, 21*, 1312-1322.

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., & Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441*, 876-479.

Frank, M.J., Doll, B.B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience, 12*, 1062-1068.

Sutton, R.S. (1990). Integrated architectures for learning, planning and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, p. 216–224, San Mateo, CA.:Morgan Kaufmann .

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

Supplemental Table 1. The number of excluded trials and the values of the potential confound variables on exploration and exploitation trials after correction.

	# excluded trials	Exploration	Exploitation	*p*-value
Expected pay-off	83.8 (13.5)	60.0 (4.1)	60.1 (4.0)	.29
Entropy	27.0 (14.0)	1.25 (0.30)	1.26 (0.29)	.02
Pay-off preceding trial	23.2 (15.9)	57.8 (2.0)	57.9 (2.0)	.13
Prediction error preceding trial	14.3 (9.6)	-1.81 (5.24)	-1.83 (5.28)	.40

Note: SD in parentheses. The difference in entropy after correction is in the opposite direction (larger entropy on exploitation trials) compared to the original effect.

Supplemental Table 2. Mean parameter estimates and negative log likelihood for the fit of the softmax model to the choice data of each participant. The parameter values used to generate the pay-offs, and the negative log likelihood of a model in which choices are made randomly are also shown.

	Estimated values	Generative values
β	0.160 (0.066)	
λ	0.894 (0.083)	0.9836
θ	56.9 (17.6)	50
σ_d	50 (fixed)	2.8
σ_0	4 (fixed)	4
-LL	153.1 (34.8)	
-LL randomly choosing model	247.2 (2.0)	

Note: SD in parentheses; -LL = negative log likelihood