ARTICLES
https://doi.org/10.1038/s41593-019-0518-9
nature neuroscience

Computational noise in reward-guided learning drives behavioral variability in volatile environments

Charles Findling[1,2,4], Vasilisa Skvortsova[1,4], Rémi Dromnelle[1,3], Stefano Palminteri[1] and Valentin Wyart[1*]

When learning the value of actions in volatile environments, humans often make seemingly irrational decisions that fail to maximize expected value. We reasoned that these 'non-greedy' decisions, instead of reflecting information seeking during choice, may be caused by computational noise in the learning of action values. Here using reinforcement learning models of behavior and multimodal neurophysiological data, we show that the majority of non-greedy decisions stem from this learning noise. The trial-to-trial variability of sequential learning steps and their impact on behavior could be predicted both by blood oxygen level-dependent responses to obtained rewards in the dorsal anterior cingulate cortex and by phasic pupillary dilation, suggestive of neuromodulatory fluctuations driven by the locus coeruleus–norepinephrine system. Together, these findings indicate that most behavioral variability, rather than reflecting human exploration, is due to the limited computational precision of reward-guided learning.

In uncertain environments, decision-makers can learn rewarding actions by trial-and-error to maximize their expected payoff (Fig. 1a). An important challenge is that reward contingencies typically change over time, and thus a less-rewarded action at a given point in time can become more rewarding later (Fig. 1b). Versatile machine learning algorithms, known collectively as reinforcement learning (RL), describe the changing values of possible actions and the policy used to choose among them[1]. One biologically plausible class of RL models updates the expected values associated with possible actions sequentially based on the prediction error between obtained and expected reward—a learning scheme known as the Rescorla–Wagner rule[2]. At any given time point, the decision-maker chooses on the basis of the difference in expected value between possible actions, by selecting the action associated with the largest expected reward.

However, in volatile environments in which reward contingencies change rapidly over time, human decision-makers make a substantial number of 'non-greedy' decisions that do not maximize the expected value predicted by reinforcement learning[3] (in contrast to value-maximizing, 'greedy' decisions). Prominent theories describe these non-greedy decisions as the result of a compromise between exploiting a currently well-valued action versus exploring other, possibly better-valued actions—known as the exploration–exploitation trade-off. In this view, information seeking motivates non-greedy decisions. Indeed, for a value-maximizing agent, lower-valued actions are chosen less often and thus their expected values are more uncertain than those of higher-valued actions. Non-greedy decisions in favor of recently unchosen actions thus reduce uncertainty about their current value and increase long-term payoff[4]. An important, implicit corollary of this view is that the underlying RL process updates action values without any internal variability after each obtained reward.

However, it has recently been shown that the accuracy of human perceptual decisions based on multiple sensory cues is bounded not by variability in the choice process, but rather by inference noise arising during the accumulation of evidence[5,6]. An intriguing possibility is that the learning process at the center of reward-guided decision-making might be subject to the same kind of computational noise, in this case random variability in the update of action values (Fig. 1c). Critically, the existence of intrinsic noise in RL would trigger non-greedy decisions owing to random deviations between exact applications of the learning rule and its noisy realizations following each obtained reward. In this view, an unknown fraction of non-greedy decisions would not result from overt information seeking during choice, as assumed by existing theories and computational models, but from the limited precision of the underlying learning process.

To determine whether, and to what extent, learning noise drives non-greedy decisions during reward-guided decision-making, we first derived a theoretical formulation of RL that allows for random noise in its core computations. In a series of behavioral and neuroimaging experiments, tested on a total of 90 human participants, we then quantified the fraction of non-greedy decisions that could be attributed to learning noise, and identified its neurophysiological substrates using functional magnetic resonance imaging (fMRI) and pupillometric recordings.

Results

Experimental protocol and computational model. We designed a restless, two-armed bandit game. In three experiments, human participants were asked to maximize their monetary payoff by sampling repeatedly from one of two reward sources depicted by colored shapes (Fig. 1a, see Methods). The payoffs that could be obtained from either shape were sampled from probability

[1]Laboratoire de Neurosciences Cognitives et Computationnelles, Inserm U960, Département d'Études Cognitives, École Normale Supérieure, PSL University, Paris, France. [2]ENSAE ParisTech, Paris-Saclay University, Palaiseau, France. [3]Institut des Systèmes Intelligents et de Robotique, CNRS UMR7222, Sorbonne University, Paris, France. [4]These authors contributed equally: Charles Findling, Vasilisa Skvortsova. *e-mail: valentin.wyart@ens.fr

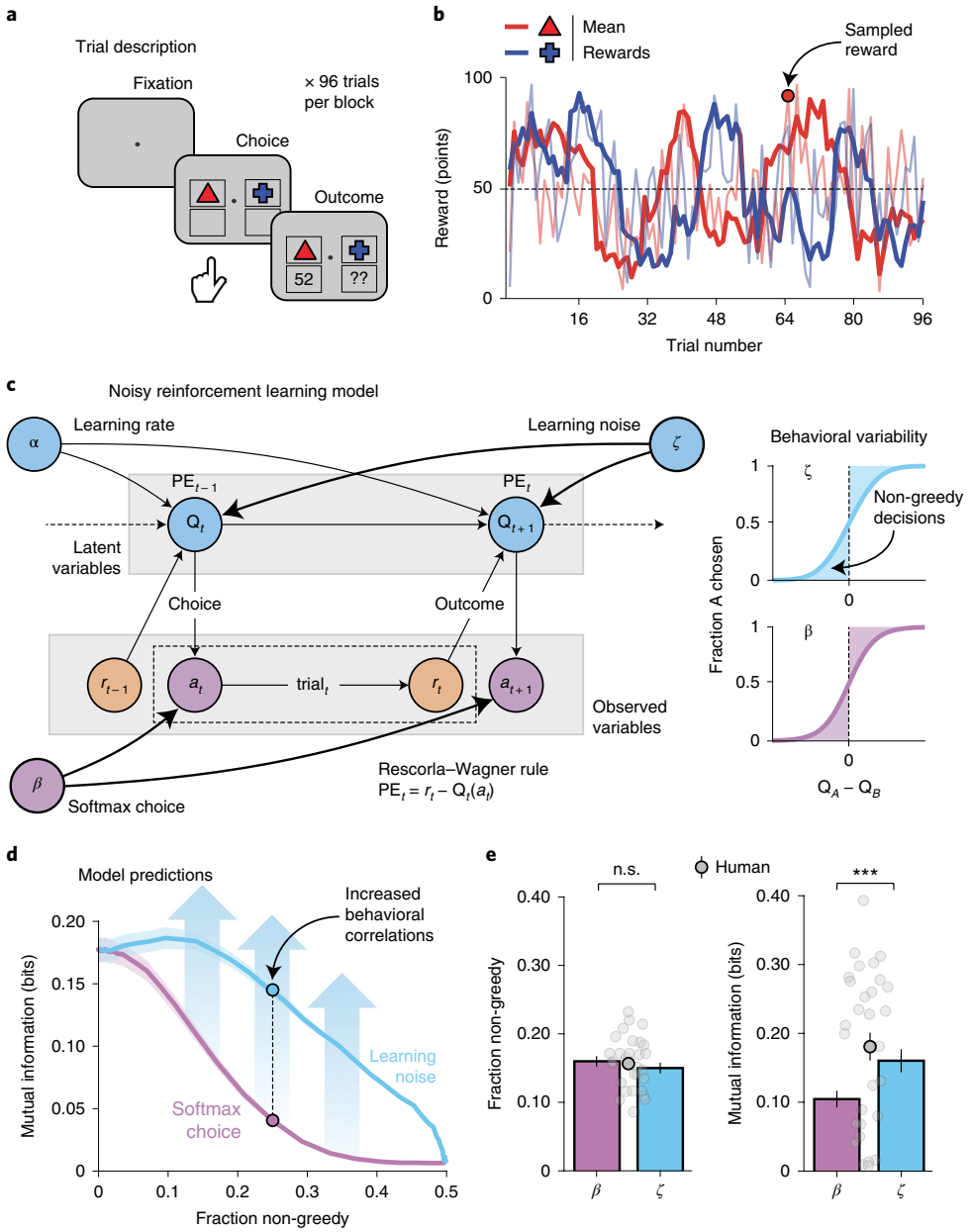


Fig. 1 | Experimental paradigm and noisy RL model. a, Trial structure in the restless, two-armed bandit task divided into short blocks of trials. In each trial, participants were asked to choose one of two reward sources depicted by colored shapes, and then observed its associated outcome (from 1 to 99 points, converted into real financial incentives at the end of the experiment). **b**, Example of drifts in the magnitude of rewards that can be obtained from the two sources. Rewards were sampled from probability distributions with means that drifted independently across trials. Thick lines represent the drifting means of the two probability distributions, whereas thin lines correspond to reward samples drawn from the probability distributions that can be obtained if chosen in each trial. **c**, Graphical representation of the noisy RL model used to fit human behavior in the task. The Rescorla–Wagner learning rule applied to update action values is corrupted by additive random noise. The choice process is modeled using a stochastic softmax action selection policy. Learning noise is assumed to be negligible in exact RL models. Right: illustration of the fraction of non-greedy decisions predicted either by an exact RL model followed by a noisy RL model followed by a purely value-maximizing action selection policy (top; area shaded in blue), or by a softmax action selection policy (bottom; area shaded in purple). Exact and noisy RL models are indistinguishable based only on their predicted fraction of non-greedy decisions. **d**, Predicted relationship between the fraction of non-greedy decisions and the mutual information of successive decisions for the exact (purple) and noisy (blue) RL models. For the same fraction of non-greedy decisions, a noisy RL model predicts larger behavioral correlations (mutual information) across successive decisions than an exact RL model. **e**, Falsification of the exact RL model through model simulations. Simulated (bars) and observed (dot) fraction of non-greedy decisions (left) and mutual information of successive decisions (right). Pale gray dots indicate individual (participant-level) observations. Although the overall fraction of non-greedy decisions is captured well by both noisy and exact RL models, the observed mutual information is predicted more accurately by simulations of noisy RL than exact RL. Error bars, s.e.m. The statistical tests performed are two-tailed (paired) *t*-tests (experiment 1, $n=29$). ***$P<0.001$. n.s., non-significant ($P>0.2$).

distributions with means that drifted independently across trials, thereby encouraging participants to track these mean values over time (Fig. 1b).

To characterize the origin of non-greedy decisions made in this task, we derived a RL model (Fig. 1c, see Methods) in which the Rescorla–Wagner rule applied to update action values is corrupted

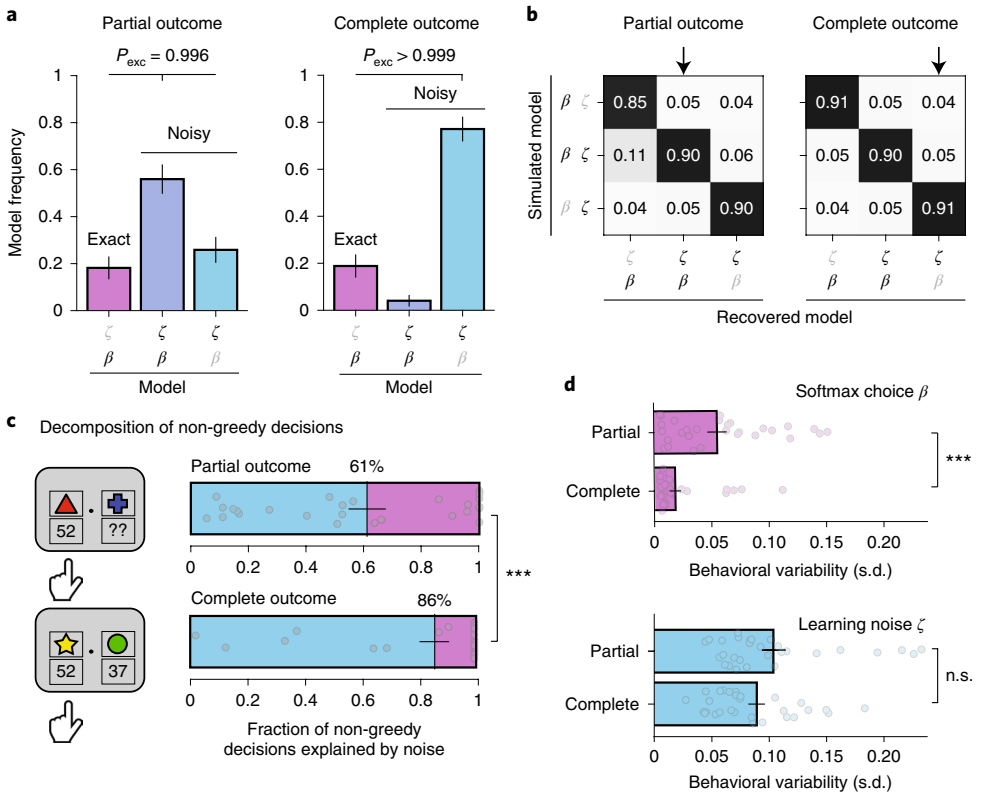


Fig. 2 | Contributions of learning noise and choice stochasticity to non-greedy decisions. a, BMS results in the partial outcome condition (left) and the complete outcome condition (right), pooled across experiment 1 and experiment 2 (total $n = 59$). Estimated model frequencies for exact RL (β, left bar), noisy RL with a softmax action selection rule (β and ζ, middle bar) and noisy RL with an argmax action selection rule (ζ, right bar). Noisy RL outperforms exact RL in both outcome conditions. Error bars, s.d. of the estimated Dirichlet distribution. P_{exc}, exceedance P value. **b**, Model recovery results in the partial outcome condition (left) and the complete outcome condition (right). Confusion matrices displaying the estimated model frequencies of exact and noisy RL (columns) obtained by simulations of exact and noisy RL (rows). Bayesian model selection enables the simulated, ground-truth RL model to be accurately recovered in both outcome conditions. **c**, Fraction of non-greedy decisions explained by learning noise (blue area) and choice stochasticity (purple area) in the partial outcome condition (top) and the complete outcome condition (bottom). Learning noise alone explains more than half of non-greedy decisions in the partial outcome condition, and almost all non-greedy decisions in the complete outcome condition. **d**, Amounts of behavioral variability (expressed as s.d. on the difference between action values predicted by the noisy RL model) due separately to learning noise (bottom) and choice stochasticity (top) in the partial and complete outcome conditions. Noise-driven variability does not differ between the two outcome conditions, whereas choice-driven variability is strongly reduced in the complete outcome condition. Error bars, s.e.m. The statistical tests performed are two-tailed (paired) t-tests (experiment 1, $n = 29$). ***$P < 0.001$. n.s., non-significant ($P > 0.2$).

by random noise with a standard deviation (s.d.) equal to a fraction (ζ) of the magnitude of the prediction error. This multiplicative structure of the learning noise follows the ubiquitous Weber's law of intensity sensation[7] prevalent in numerous perceptual domains (including vision, numerosity and time) and in the magnitude of associated neural responses[8,9]. As in existing theories, we modeled the choice process using a stochastic 'softmax' action selection policy, controlled by an inverse temperature, β. Importantly, the two sources of behavioral variability make different predictions regarding the temporal structure of decisions across successive trials. Indeed, learning noise corrupts the action values that are gradually updated across trials. By contrast, choice stochasticity reads out action values without altering them and is independently distributed across trials. Therefore, for the same fraction of non-greedy decisions simulated using either learning noise or choice stochasticity, learning noise engenders larger behavioral correlations across successive decisions (Fig. 1d).

Dominant contribution of learning noise to non-greedy decisions. In the first neuroimaging experiment (experiment 1, $n = 29$), participants selected the shape associated with the largest true mean in the majority of trials (mean ± s.e.m., 64.9 ± 0.9%; t-test against chance, $t_{28} = 17.2$, $P < 0.001$). As anticipated, a substantial fraction of decisions made were non-greedy decisions, which do not maximize expected value with respect to an exact (noise-free) RL model (15.7 ± 0.7%).

We performed Bayesian model selection (BMS) to quantify the contributions of learning- and choice-driven sources of variability to non-greedy decisions (Fig. 2a). Using particle filtering procedures to estimate the model evidence conditioned on human decisions (see Methods and Supplementary Modeling), we found that the RL model featuring both learning noise and a softmax action selection policy explained human behavior significantly better than RL models featuring either of these two sources of behavioral variability (exceedance $P = 0.996$, see Supplementary Fig. 1 for a parameter knock-out procedure). To validate these findings, we implemented a model recovery procedure[10], which confirmed that our BMS procedure was capable of correctly distinguishing learning noise from choice stochasticity in our task (Fig. 2b, see Supplementary Fig. 2 for a parameter recovery procedure). We could also empirically falsify[10] the exact RL model (Fig. 1e, see Methods): although both exact and noisy RL models fitted to human behavior accounted for the observed fraction of non-greedy decisions, the exact RL model predicted a lower mutual information between successive decisions

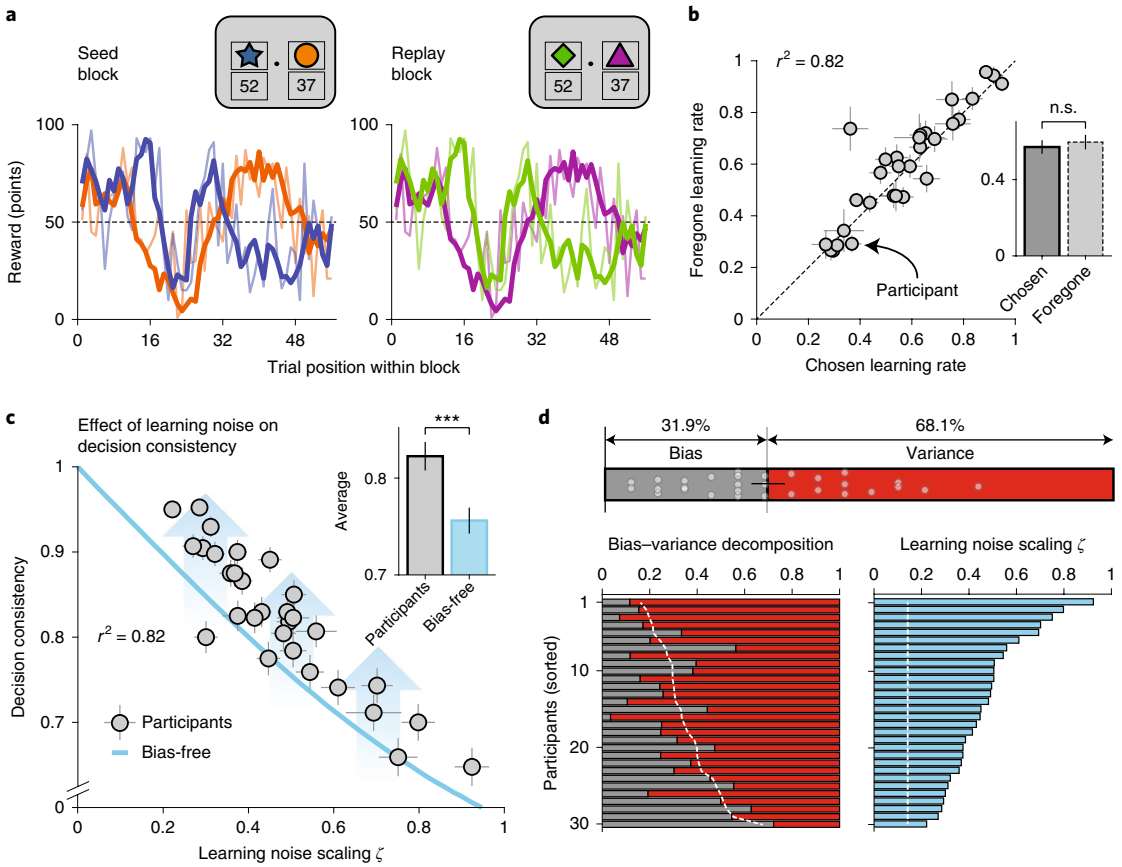


Fig. 3 | Decomposition of learning noise into ultimately predictable and unpredictable terms. a, Trial structure in the experiment consisting of repeated blocks of the same sequence of rewards, tested only in the complete outcome condition. Unknown to participants, each seed block of trials (left) was replayed at another point in the experiment (right) using different colored shapes. Colored lines correspond to the mean rewards associated with each symbol. **b**, Learning rates associated with chosen and foregone actions (dots ± error bars, means ± s.d.). Learning rates did not differ between chosen and foregone actions, indicating that action values in a given trial do not depend on choices made in earlier trials ($t_{29} = 1.6$, $P = 0.123$). **c**, Decision consistency (y axis)—that is, the fraction of trials in which the same decision was made in seed and replay blocks—as a function of the scaling ζ of learning noise with the magnitude of learning steps (x axis) fitted by the noisy RL model to each participant. The blue line corresponds to the predicted relationship between the two quantities through simulations of the noisy RL model. Variability across tested participants (dots ± error bars, means ± s.d.) shows the negative correlation predicted by simulations of the noisy RL model. Decision consistency nevertheless exceeds predictions from the noisy RL model, as indicated by the blue arrows. **d**, Bias–variance decomposition of learning noise. Top: fraction of learning noise explained by predictable biases (gray area) and unpredictable variance (red area). More than two-thirds of learning noise was not attributable to systematic deviations from the Rescorla–Wagner rule used in our noisy RL model. Bottom: relationship between the Weber scaling ζ of learning noise (right) and the bias–variance decomposition of learning noise (left) across participants. Sorting participants by ζ reveals a clear increase in the variance term with ζ, indicating that ζ indexes random variability in learning rather than systematic deviations from the Rescorla–Wagner rule. Error bars, s.e.m. (unless indicated otherwise). The statistical tests performed are two-tailed (paired) t-tests (experiment 3, $n = 30$). ***$P < 0.001$. n.s., non-significant (see **b** for the precise P value).

(exact, 0.073 ± 0.008 bit; noisy, 0.111 ± 0.011 bit; paired t-test, $t_{28} = 5.8$, $P < 0.001$), substantially lower than the mutual information between successive human decisions (0.126 ± 0.016 bit).

Given the presence of both sources of behavioral variability, we quantified the precise contributions of learning noise and choice stochasticity to non-greedy decisions. For this purpose, we first estimated the trial-to-trial trajectories of latent action values corrupted by learning noise conditioned on observed human decisions in each block (see Methods). We verified that trial-to-trial realizations of learning noise conformed to their distributional assumptions (see Supplementary Modeling). We then assessed the fraction of non-greedy decisions for which noisy realizations of the learning rule resulted in an opposite ranking of action values to exact applications of the same rule. This quantitative analysis revealed that learning noise alone explained as much as 60.6 ± 6.6% of non-greedy decisions (Fig. 2c). This pattern of findings, fully replicated in an additional behavioral experiment (experiment 2, $n = 30$, see Supplementary Modeling), indicates that behavioral variability is driven to a large extent by random noise in the update of action values, rather than by stochasticity in the choice process. This pattern of findings is robust to alternative definitions of non-greedy decisions, in particular based on the optimal model for learning action values in our task (Kalman filtering, see Supplementary Modeling).

Dissociating learning noise from information seeking. We then sought to dissociate the observed learning noise from information seeking. One obvious way of achieving this consists of showing that the behavioral variability stemming from learning noise is not aimed explicitly at seeking information about recently unchosen actions, the associated rewards of which have not been observed and are thus uncertain. To test this important prediction, we contrasted in both experiments the classical 'partial outcome' condition, in which participants observe only the reward yielded by the selected shape (Fig. 2c), with another 'complete outcome' condition,

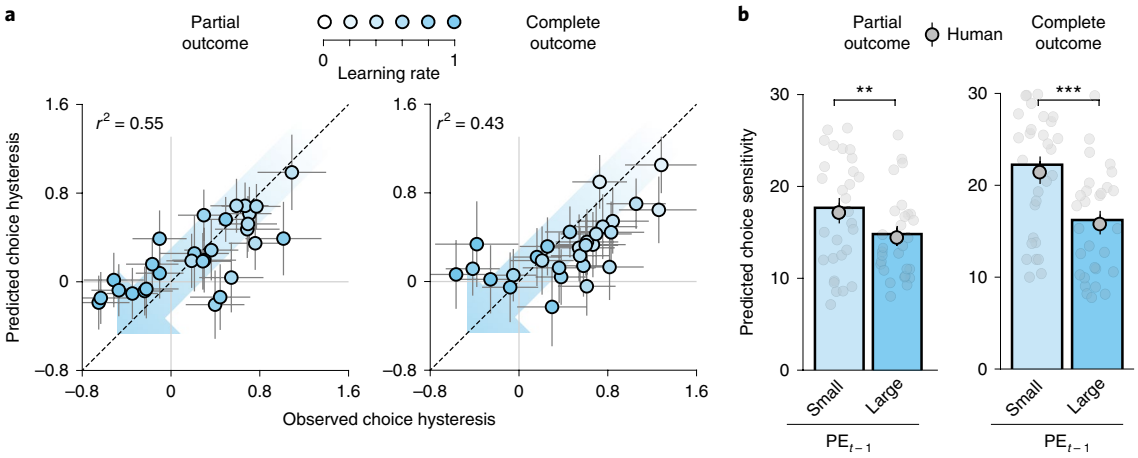


Fig. 4 | Characterization of decision effects predicted by learning noise. a, Relationship between the choice hysteresis measured in tested participants (*x* axis) and the choice hysteresis predicted by simulations of the noisy RL model (*y* axis) in the partial outcome condition (left) and the complete outcome condition (right). Dots ± error bars, mean ± s.d. of estimated posterior distributions for each tested participant. The blue shading of the dots indicates the learning rate associated with the chosen action (higher saturation indicates faster learning rates). The choice hysteresis measured in tested participants correlates with the apparent choice hysteresis predicted by simulations of the noisy RL model in both conditions. Choice hysteresis also decreases with learning rate (shaded arrow). **b**, Adaptation of choice stochasticity to surprise in tested participants (dots) and simulations of the noisy RL model (bars) in the partial outcome condition (left) and the complete outcome condition (right). The sensitivity of human choices fitted using an exact RL model (and measured through its best-fitting inverse temperature β) is significantly lower in trials following larger-than-average prediction errors, in both partial and complete outcome conditions. Simulations of the noisy RL model exhibit the same adaptation of choice sensitivity to surprise. Error bars, s.e.m. The statistical tests performed are two-tailed (paired) *t*-tests (experiment 1, $n = 29$). $^{**}P < 0.01$. $^{***}P < 0.001$.

in which participants additionally observe the foregone reward that would have been obtained if the other, unchosen shape had been selected[11,12] (Fig. 2c). In this additional condition, performed by the same participants, there is by definition no incentive to explore (that is, to choose actions that do not maximize expected value), given that there is equal uncertainty about the values of chosen and unchosen actions.

Interestingly, participants made a lower, but still substantial, proportion of non-greedy decisions in the complete outcome condition (partial, 15.7 ± 0.7%; complete, 11.9 ± 0.7%; paired *t*-test, $t_{28} = -4.2$, $P < 0.001$). BMS showed that the noisy RL model featuring a purely value-maximizing 'argmax' policy best explained human behavior (Fig. 2a, exceedance $P > 0.999$, see Supplementary Fig. 1 for a parameter knock-out procedure). Consequently, and in contrast to what was observed in the partial outcome condition, learning noise explained almost all non-greedy decisions in this condition (86.1 ± 5.2%, complete versus partial, paired *t*-test, $t_{28} = 3.6$, $P = 0.001$; Fig. 2c).

We further confirmed that the increased fraction of non-greedy decisions explained by learning noise is due to lower choice stochasticity rather than to increased learning noise. Instead of computing the relative fraction of non-greedy decisions driven by learning noise, we estimated the absolute amount of behavioral variability due to learning noise and to choice stochasticity separately (Fig. 2d, see Methods). This analysis confirmed that choice-driven variability was reduced in the complete outcome condition (partial, 0.078 ± 0.010; complete, 0.025 ± 0.007; paired *t*-test, $t_{28} = -4.6$, $P < 0.001$), whereas noise-driven variability did not differ between the two conditions (partial, 0.110 ± 0.010; complete, 0.093 ± 0.007; paired *t*-test, $t_{28} = -1.2$, $P = 0.240$, Bayes factor quantifying the evidence in favor of the null hypothesis (BF_{H0}) = 2.6). Together, these findings indicate that learning noise does not aim at seeking information about recently unchosen actions. This pattern of findings was replicated in experiment 2 using the same experimental design (see Supplementary Modeling).

Dissociating learning noise from model misspecification. One important possible confounding factor is that part of the observed learning noise would be caused not by random deviations around the proposed Rescorla–Wagner rule, but by systematic deviations from this canonical learning rule (in other words, a misspecification of our learning model). To decompose learning noise into random and systematic deviations, we ran a third, behavioral experiment (experiment 3, $n = 30$) in which we estimated the consistency of human decisions across repetitions of the same sequence of rewards—that is, the fraction of trials in which the same decision was made in the two repeated blocks (Fig. 3a, see Methods).

We restricted this experiment to the complete outcome condition for two important reasons. First, we wanted behavioral variability to be driven solely by learning noise and not by a softmax action selection policy—a result replicated in this additional experiment (exceedance $P = 0.997$). Second, we wanted the action values predicted by an exact RL model to be identical across the two repetitions of the same block, irrespective of the decisions made in the two repeated blocks. This was ensured by the observation that participants learned equally from chosen and foregone rewards (Fig. 3b, learning rate α, chosen: 0.571 ± 0.036, foregone: 0.597 ± 0.038; paired *t*-test, $t_{29} = 1.6$, $P = 0.123$, $BF_{H0} = 1.7$).

We applied a recently developed statistical approach[6] (see Methods) to split the overall behavioral variability into a predictable bias term (reflecting systematic deviations from the Rescorla–Wagner rule) and an unpredictable variance term (reflecting random deviations around this learning rule). In practice, the consistency of decisions across repeated blocks, which ranged from 64.8% to 95.2% across participants (mean ± s.e.m., 82.3 ± 1.5%) and was essentially constant over the course of repeated blocks (see Supplementary Modeling), was used to estimate the bias–variance split of learning noise. Indeed, systematic deviations tend to increase the consistency of decisions across repeated blocks, whereas random deviations tend to decrease self-consistency. Therefore, participants with more learning noise should be less consistent across repeated blocks if learning noise reflects random deviations, but not if learning noise reflects systematic deviations (see Supplementary Modeling for simulations of misspecified models). We first

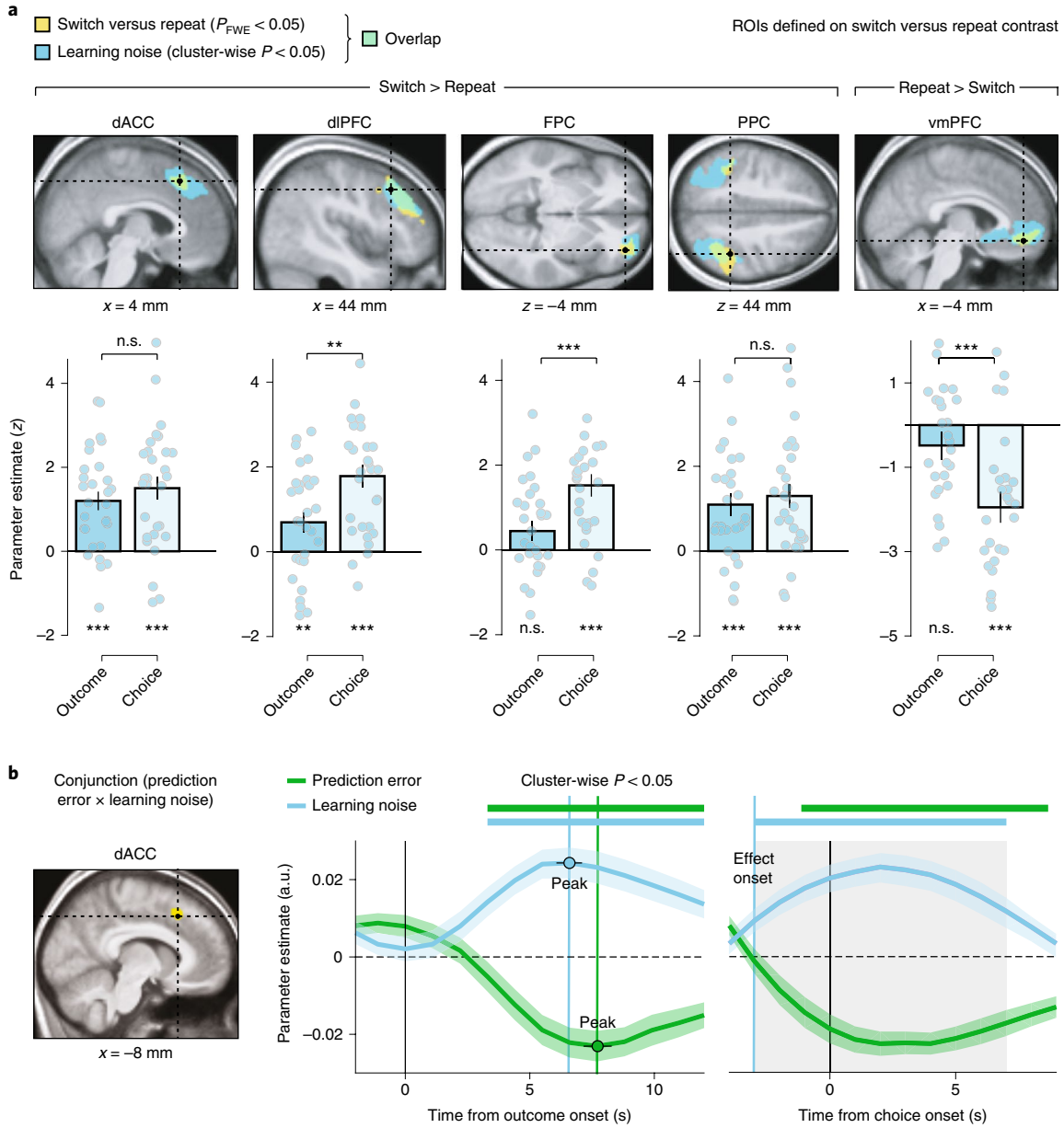


Fig. 5 | Neural correlates of learning noise in the human brain. a, Top: ROIs (yellow) obtained for the switch minus repeat contrast corrected at a whole-brain family-wise error rate (FWE) of 0.05. Regions shaded in blue indicate the clusters that correlate significantly with learning noise at a cluster-wise corrected *P* value of 0.05. Neural correlates of learning noise overlap broadly with neural correlates of the switch minus repeat contrast. Bottom: group-level parameter estimates for the learning noise regressor in each of the ROIs defined by the switch minus repeat contrast locked to the onset of the outcome period of trial $t-1$ (left bar) and the choice period of trial t. Error bars, s.e.m. The statistical tests performed are two-tailed (paired) *t*-tests (experiment 1, $n=29$). $^{**}P<0.01$. $^{***}P<0.001$. n.s., non-significant ($P>0.2$). **b**, Left: results of the whole-brain conjunction analysis of prediction error and learning noise at a cluster-wise corrected *P* value of 0.05. Only the dACC reflects simultaneously the learning signal (that is, the prediction error associated with the chosen action) and its trial-to-trial learning noise. Right: results of the finite impulse response analysis showing the temporal dynamics of prediction error and learning noise in dACC activity, locked to outcome presentation (left) and to the onset of the following choice period (right). The correlations between dACC activity and prediction error (green) and learning noise (blue) peak at approximately the same time following outcome presentation (dots ± error bars, jackknifed means ± s.e.m. for correlation peaks estimated separately for the two regressors). Thick horizontal lines indicate time windows in which parameter estimates diverge significantly from zero (cluster definition threshold $P=0.05$, cluster-wise $P<0.001$). The correlation between dACC and learning noise emerges significantly before the onset of the following choice period (jackknifed $t_{28}=-2.7$, $P=0.012$). Brain coordinates are expressed in Montreal Neurological Institute (MNI) coordinate space.

observed that participants with greater learning noise showed lower decision consistency across repeated blocks (Fig. 3c, linear correlation, $r^2=0.829$, d.f. = 28, $P<0.001$). This supports our proposal that most of the learning noise captured by the model is due to random variability rather than to systematic deviations from the Rescorla–Wagner rule. Consistently, the split that best accounted for the consistency of human decisions was 31.8 ± 3.2% for the bias term and 68.2 ± 3.2% for the variance term (Fig. 3d). This result indicates that two-thirds of learning noise is not attributable to misspecifications of our model and supports our proposal that learning

noise primarily reflects the limited computational precision of reward-guided learning.

Explaining decision effects as consequences of learning noise. During sequential learning in volatile environments, humans make non-greedy decisions in favor of recently unchosen actions, but also show a tendency to repeat their previous decision over and above the difference between action values. This decision effect has been described in computational terms by an additional bias in the choice process, often termed choice hysteresis, which could have beneficial (stabilizing) properties for the decision-maker[13–15]. We realized that this effect falls naturally out of the statistical properties of learning noise, owing to intrinsic temporal correlations in the noise-corrupted action values used by the decision-maker. Using an exact RL model to fit human decisions, we observed a positive choice hysteresis in both partial and complete outcome conditions (*t*-test against zero: partial, $t_{28}=2.6$, $P=0.013$; complete, $t_{28}=5.0$, $P<0.001$). Critically, the choice hysteresis measured in participants correlated with the apparent choice hysteresis predicted by the noisy RL model (Fig. 4a, linear correlation: partial, $r^2=0.553$, d.f.$=27$, $P<0.001$; complete, $r^2=0.425$, d.f.$=27$, $P<0.001$). Furthermore, adding explicit choice hysteresis to the exact RL model provided a significantly worse account of human behavior than the noisy RL model in both outcome conditions (partial, exceedance $P=0.040$; complete, exceedance $P<0.001$). This finding supports our proposal that choice hysteresis is not caused by an explicit bias in the choice process, but rather by learning noise that propagates through corrupted action values across successive decisions.

A second effect documented in the literature consists of an adjustment of choice stochasticity to surprise (that is, the magnitude of the prediction error in RL)[16,17]. Like choice hysteresis, this decision effect falls naturally out of learning noise, owing to its scaling with the magnitude of the prediction error. Using an exact RL model to fit human decisions, we observed a decreased choice sensitivity to action values (computed using a logistic regression of choice against model-predicted action values, see Methods) in trials following larger-than-average prediction errors (Fig. 4b) in both partial and complete outcome conditions (paired *t*-test: partial, $t_{28}=-2.9$, $P=0.007$; complete, $t_{28}=-3.7$, $P<0.001$). Simulations of the noisy RL model fitted using an exact RL model featured the same adaptation to surprise (paired *t*-test: partial, $t_{28}=-5.5$, $P=0.001$; complete, $t_{28}=-8.9$, $P<0.001$). Importantly, the adaptation predicted by simulations of the noisy RL model matched both the direction and the size of the adjustment observed in participants (paired *t*-test: partial, $t_{28}=0.2$, $P=0.847$, $BF_{H0}=5.0$; complete, $t_{28}=0.3$, $P=0.778$, $BF_{H0}=4.9$). Like choice hysteresis, adding an explicit adjustment of choice stochasticity to surprise provided a significantly worse account of human behavior than the noisy RL model (see Supplementary Modeling). These results suggest that the adaptation of choice stochasticity to surprise is caused by the multiplicative structure of learning noise, rather than by overt information seeking following surprising outcomes.

Neural correlates of learning noise in the frontal cortex. To identify the neural mechanisms underlying this undocumented learning noise, we analyzed blood oxygen level-dependent (BOLD) fMRI data (experiment 1, $n=29$) recorded while participants performed the task (see Methods). As documented in the literature, a classical network of brain regions implicated in cognitive control showed increased BOLD responses to switches away from the previous action[18,19] (Fig. 5a and Supplementary Table 1), including the dorsal anterior cingulate cortex (dACC), the dorsolateral prefrontal cortex (dlPFC), the frontopolar cortex (FPC) and the posterior parietal cortex (PPC).

Locked to outcome presentation, an overlapping brain network positively reflected the magnitude of learning noise—defined as trial-to-trial deviations from exact applications of the Rescorla–Wagner rule (Fig. 5a, see Methods and Supplementary Modeling), including the dACC, the right dlPFC and the PPC (cluster-corrected $P<0.05$, Supplementary Table 2). Importantly, a conjunction analysis revealed that, among these three brain regions, only the dACC simultaneously reflected the magnitude of learning noise and the prediction error associated with the obtained reward (cluster-corrected $P<0.05$, Fig. 5b). To characterize the temporal dynamics of learning correlates in dACC activity, we constructed a finite impulse response model aligned either to the presentation of each outcome or to the presentation of the following choice (see Supplementary Modeling). The correlation of dACC activity with learning noise and prediction errors followed similar time profiles after outcome presentation (Fig. 5b). In the model aligned to the presentation of the following choice, the correlation of dACC activity with learning noise emerged significantly before choice onset (jackknifed $t_{28}=-2.7$, $P=0.012$). Together, these results indicate that dACC responses to obtained rewards reflect the mean and variability of learning steps during the update of action values.

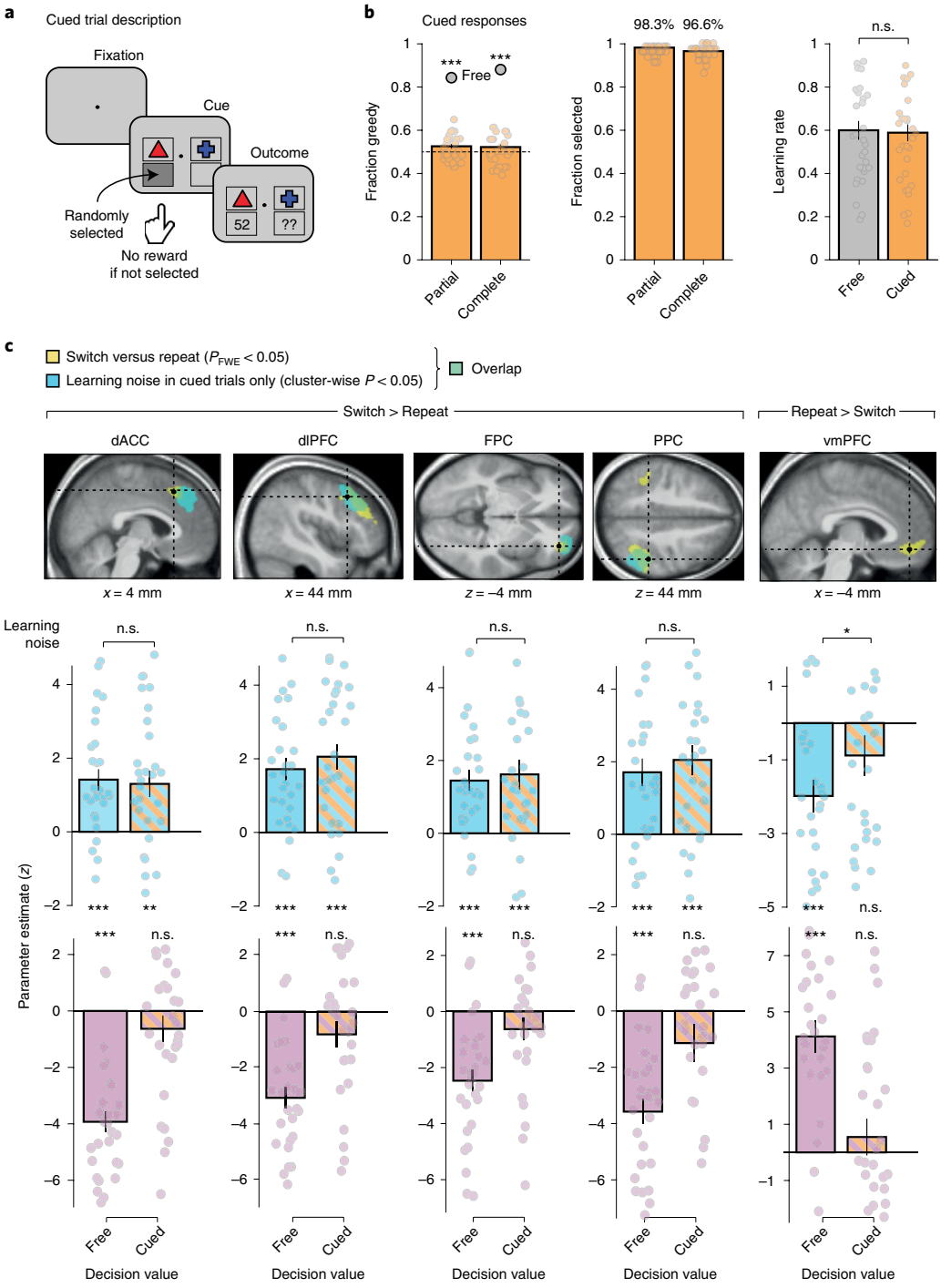
During the following choice period when learning noise translates into behavioral variability, the magnitude of learning noise was again reflected positively in the dACC and the right dlPFC, but was also reflected positively in the FPC and negatively in the ventromedial prefrontal cortex (vmPFC) at a conservative statistical threshold (family-wise error-corrected $P<0.05$, Fig. 5a). Importantly, the dACC reflected learning noise equally strongly in the outcome and following choice periods (dACC, $t_{28}=0.8$, $P=0.390$, $BF_{H0}=3.6$). By contrast, the right dlPFC, the FPC and the vmPFC reflected learning noise more strongly in the choice period than in the preceding outcome period (dlPFC $t_{28}=2.9$, $P=0.008$; FPC $t_{28}=3.1$,

Fig. 6 | Neural correlates of learning noise in choice-free, cued trials. a, Trial structure in cued trials (one-quarter of all trials). In cued trials, participants were required to select the highlighted action that was randomly pre-selected, and then observed its associated outcome (and the foregone outcome in the complete outcome condition) as in standard, free trials. **b**, Human behavior and learning in cued trials. Left: fraction of greedy (value-maximizing) actions in cued trials. As instructed, participants did not choose the highest-valued action in cued trials (dots indicate the average fraction of greedy actions found in free trials as reference) in either the partial outcome condition (left bar) or the complete outcome condition (right bar). Middle: fraction of actions matching the pre-selected shape in cued trials. As instructed, participants almost invariably selected the highlighted action in both the partial outcome condition (left bar) and the complete outcome condition (right bar). Right: learning rates associated with the selected action estimated in free trials (left bar) and cued trials (right bar). Learning rates did not differ between free and cued trials, indicating that participants learned equally from the two types of trials. **c**, Top: in contrast to Fig. 5, regions shaded in blue indicate the clusters that correlate significantly with learning noise in cued trials at a cluster-wise corrected *P* value of 0.05. Neural correlates of learning noise remain significant in cued trials in all ROIs except the vmPFC. Bottom: group-level parameter estimates for the learning noise regressor (above) and the decision value regressor (below) in each of the ROIs defined by the switch minus repeat in free trials (left bar) and cued trials (right bar). The decision value regressor is defined as the value difference between selected and unselected actions. The correlation of BOLD activity with decision value is significantly reduced in cued trials in all ROIs, whereas the correlation with learning noise is unchanged in all ROIs except the vmPFC. Error bars, s.e.m. The statistical tests performed are two-tailed (paired) *t*-tests (experiment 1, $n=29$). $^{*}P<0.05$. $^{**}P<0.01$. $^{***}P<0.001$. n.s., non-significant ($P>0.2$ unless indicated otherwise).

$P = 0.005$; vmPFC $t_{28} = -3.3$, $P = 0.003$). Together, this pattern of findings assigns a unique position to the dACC among the neural correlates of learning noise: in addition to its correlation with the mean and variability of learning steps during the update of action values, dACC activity reflects learning noise with the same intensity during the following choice period when learning noise triggers behavioral variability.

Dissociating neural correlates of learning noise from choice. To rule out the possibility that what is captured as learning noise arises from a non-modeled property of the choice process, experiment 1 included not only choice trials, in which subjects could select the option they wanted to sample, but also cued trials, in which subjects were required to select one of the two shapes (Fig. 6a, see Methods). In these cued trials, there is by definition no choice to be made, and indeed participants invariably selected the cued shape (Fig. 6b, partial, 98.3 ± 0.3%; complete, 96.6 ± 0.5%). Nevertheless, BMS revealed that participants learned equally from obtained rewards in choice and cued trials ($\alpha_{cued} = \alpha_{choice}$ versus $\alpha_{cued} = 0$, partial, BF ≈ $10^{208.6}$, exceedance $P > 0.999$; complete, BF ≈ $10^{223.4}$, exceedance $P > 0.999$), and that learning is corrupted by the same noise in cued trials as in choice trials ($\zeta_{cued} = \zeta_{choice}$ versus $\zeta_{cued} = 0$, partial, BF ≈ $10^{8.3}$, exceedance $P > 0.999$; complete, BF ≈ $10^{10.5}$, exceedance $P > 0.999$).

We then tested whether the neural correlates of learning noise found during the choice period were also present in cued trials (Fig. 6c, see Supplementary Modeling). The positive correlation



between BOLD activity and learning noise remained highly significant and unchanged in cued trials in the dACC (cued, $t_{28}=3.6$, $P=0.001$; cued versus choice, $t_{28}=-0.3$, $P=0.760$, $BF_{H0}=4.8$), the right dlPFC (cued $t_{28}=6.0$, $P<0.001$; cued versus choice $t_{28}=0.8$, $P=0.404$, $BF_{H0}=3.7$) and the FPC (cued $t_{28}=4.0$, $P<0.001$; cued versus choice, $t_{28}=-0.4$, $P=0.677$, $BF_{H0}=4.7$). This finding further strengthens our proposal that the learning noise fitted by our noisy RL model reflects variability in the update of action values, rather than an unknown property of the choice process.

Relating neural correlates of learning noise to behavioral variability. Our neuroimaging results so far indicate that BOLD activity in several brain regions (the dACC, the right dlPFC, the PPC, the FPC and the vmPFC) reflects learning noise. An important question therefore arises as to whether these different brain regions differ in their relationship with the resulting behavioral variability. To address this important question, we formulated a brain–behavior analysis to predict behavioral variability on the basis of trial-to-trial BOLD fluctuations in these five regions of interest (ROIs) (see Methods). We reasoned that a neural signal reflecting learning noise should decrease the sensitivity of participants to action values and influence the decisions of participants in the complete outcome condition, in which the choice process does not generate any behavioral variability. A fully factorial analysis (Fig. 7a), validated through model recovery (Fig. 7b), identified two active ROIs: the dACC (family-wise $P>0.999$) and the vmPFC (family-wise $P=0.630$).

Computing parameter estimates for the model including the dACC and the vmPFC in the partial outcome condition (Fig. 7c) revealed a negative effect of dACC fluctuations on sensitivity ($\beta=-0.185\pm0.026$, $t_{28}=-7.2$, $P<0.001$), and a positive effect of vmPFC fluctuations on the same metric ($\beta=0.062\pm0.026$, $t_{28}=2.4$, $P=0.023$). The effect of the dACC was substantially larger in absolute magnitude than that of the vmPFC ($t_{28}=4.2$, $P<0.001$), and conformed to the first signature of a neural signal reflecting learning noise (Fig. 7d). Furthermore, the modulation of sensitivity by dACC responses was also negative in the complete outcome condition (Fig. 7c, $\beta=-0.170\pm0.023$, $t_{28}=-7.4$, $P<0.001$). Importantly, the effects observed in the two conditions were similar ($t_{28}=0.5$, $P=0.635$, $BF_{H0}=4.6$), in line with the second signature of a neural signal reflecting learning noise. Together, these results support our proposal that dACC fluctuations reflect learning noise (which was present in both conditions) rather than choice stochasticity (which was negligible in the complete outcome condition).

Finally, we tested whether the negative modulation of sensitivity by dACC responses could be caused not by random fluctuations of action values (as predicted by learning noise), but rather by directed fluctuations in the value of switching away from the previous action (as predicted by adjustments of the exploration–exploitation trade-off). A factorial analysis including these two effect types (see Methods) revealed a selective effect of dACC responses on sensitivity (family-wise $P>0.999$), without any measurable effect on the value of switching (family-wise $P<0.001$). This selective brain–behavior relationship indicates that trial-to-trial fluctuations in dACC activity reflect learning noise rather than adjustments of the exploration–exploitation trade-off.

Pupil-linked neuromodulatory correlates of learning noise. In addition to frontal cortical contributions to non-greedy decisions, past research has identified the locus coeruleus–norepinephrine (LC–NE) system as a reliable neurophysiological correlate of behavioral variability[20]. Large phasic responses of LC neurons are associated with task disengagement and non-greedy decisions[21,22]. Although existing theories describe these effects as adjustments of the exploration–exploitation trade-off[23], we postulated that trial-to-trial fluctuations in the computational precision of update steps, reflected by dACC activity, could be mediated by neuromodulatory fluctuations driven by the LC–NE system. Because LC activity is notoriously difficult to measure by fMRI, we took advantage of the strong, known correlation between LC activity and phasic pupil dilation[24] recorded in experiment 2 ($n=24$ participants with clean pupillometric data).

In line with the literature, we observed that a switch away from the previous action was associated with larger pupillary dilation than repeating the previous action (Fig. 7e; from −2.0 to 2.9 s following choice presentation, cluster-corrected $P<0.001$). Pupillary dilation in the same time window correlated positively with the magnitude of learning noise corrupting the preceding update step (Fig. 7e, from −2.0 to 2.2 s following choice presentation, cluster-corrected $P<0.001$). Like dACC responses, pupillary dilation correlated significantly with learning noise well before choice onset (*t*-test against zero, jackknifed $t_{23}=-11.1$, $P<0.001$).

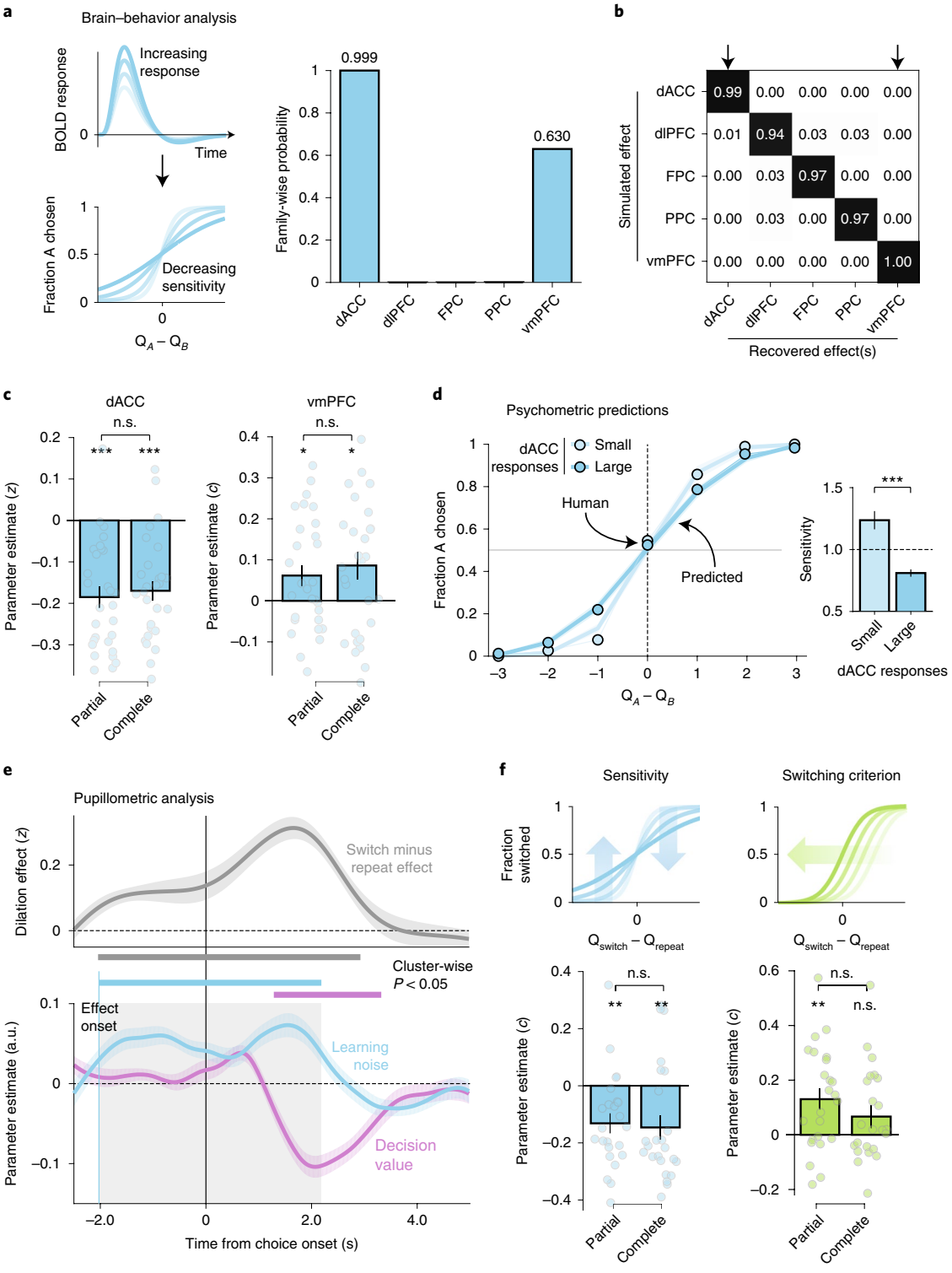
We then tested the relationship between trial-to-trial pupillary fluctuations and behavioral variability using the brain–behavior analysis previously applied to dACC responses (Fig. 7f, see Supplementary Modeling). This analysis indicated that pupillary dilation predicts both random fluctuations in sensitivity and directed

Fig. 7 | Brain–behavior and pupillometric analyses. a, Left: schematic illustration of the brain–behavior relationship predicted by a neural correlate of learning noise. Trial-to-trial variability in the amplitude of BOLD responses (shaded in blue) should correlate negatively with trial-to-trial variability in the sensitivity to action values predicted by exact application of the Rescorla–Wagner rule in the previous trial. Right: results of the full factorial brain–behavior analysis including the five ROIs identified in Fig. 5. Family-wise probability is defined as the probability that each ROI modulates sensitivity independently of the involvement of other ROIs. Only the dACC and the vmPFC have family-wise probabilities exceeding 1%. **b**, Model recovery results for the full factorial brain–behavior analysis. Confusion matrix displaying the estimated family-wise probabilities (columns) obtained for simulations of selective (single ROI) sensitivity modulations (rows). Black arrows indicate the two ROIs (dACC and vmPFC) with detected brain–behavior relationships in the data. **c**, Participant-level parameter estimates for the winning model including the dACC (left) and the vmPFC (right) in the partial outcome condition (left bars) and the complete outcome condition (right bars). Sensitivity to action values correlates negatively with dACC activity and positively with vmPFC activity in both outcome conditions. Pale blue dots indicate individual (participant-level) estimates. **d**, Psychometric brain–behavior predictions for the dACC ROI. Predicted (lines) and human (dots) psychometric curves for small (first tercile) and large (third tercile) dACC responses. Both predicted and human curves show a decreased sensitivity to action values for larger dACC responses. Inset: sensitivity estimates for small (left bar) and large (right bar) dACC responses. Bars ± error bars, jackknifed means ± s.e.m. **e**, Top: results of the switch minus repeat contrast for pupillary dilation. Pupillary dilation increases significantly before switches. Bottom: results of the finite impulse response analysis showing the temporal dynamics of learning noise (blue) and decision value (purple) in pupillary dilation, locked to the onset of the choice period. The correlation between pupillary dilation and learning noise emerges significantly before the onset of the choice period. Thick horizontal lines indicate time windows in which parameter estimates diverge significantly from zero at a temporal cluster-wise corrected *P* value of 0.01. **f**, Participant-level parameter estimates for pupil-linked modulations of sensitivity (left) and switching value or criterion (right) in the partial outcome condition (left bars) and the complete outcome condition (right bars). Pupillary dilation correlates negatively with sensitivity and positively with switching value, without a significant difference between conditions (sensitivity, $t_{28}=0.3$, $P=0.791$; switching value, $t_{28}=1.5$, $P=0.152$). Error bars, s.e.m. The statistical tests performed are two-tailed (paired) *t*-tests (experiment 1, $n=29$). $^{*}P<0.05$. $^{**}P<0.01$. $^{***}P<0.001$. n.s., non-significant ($P>0.2$ unless indicated otherwise).

fluctuations in the value of switching in the partial outcome condition (random, family-wise $P > 0.999$, $\beta = -0.132 \pm 0.035$, $t_{23} = -3.8$, $P = 0.001$; directed, family-wise $P = 0.995$, $\beta = 0.130 \pm 0.038$, $t_{23} = 3.4$, $P = 0.002$). By contrast, in the complete outcome condition, pupillary dilation predicts random fluctuations in sensitivity (family-wise $P > 0.999$, $\beta = -0.146 \pm 0.043$, $t_{23} = -3.4$, $P = 0.002$), but no directed fluctuations in the value of switching (family-wise $P = 0.095$, $\beta = 0.067 \pm 0.042$, $t_{23} = 1.6$, $P = 0.129$). Together, these results indicate that pupillary fluctuations reflect learning noise over and above adjustments of the exploration–exploitation trade-off.

Discussion

Maximizing rewards in volatile environments requires an agent to trade the exploitation of currently best valued actions against the exploration of recently unchosen, and thus more uncertain, ones. Here we sought to contrast existing information-seeking accounts with another possible source of behavioral variability: the limited computational precision of the learning process, which updates the expected values of possible actions following each reward. By decomposing behavioral variability into these two components using our noisy RL model, we show that more than half of



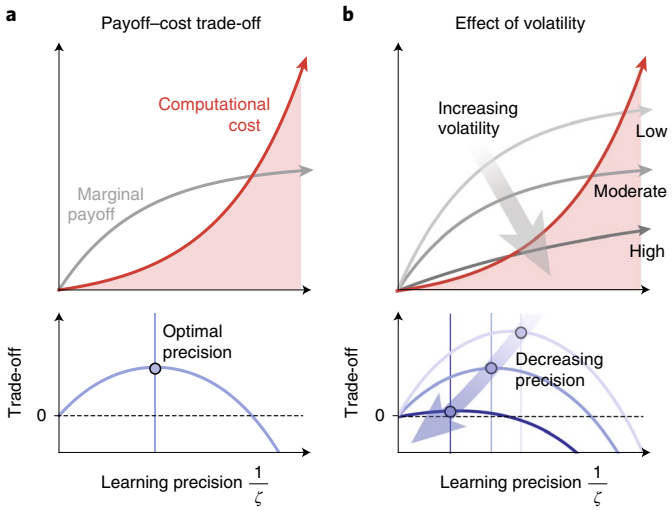


Fig. 8 | Proposed payoff–cost trade-off on learning precision. a, Illustration of the proposed payoff–cost trade-off. Top: schematic illustration of the dependencies between learning precision $1/\zeta$ (x axis) and the marginal payoff of learning (gray curve) and the computational cost of learning (red curve). The marginal payoff of learning saturates at high precisions, whereas its computational cost grows exponentially. Bottom: payoff minus cost trade-off showing an optimal learning precision that maximizes the difference between payoff and cost. **b**, Illustration of the effect of volatility on the proposed payoff–cost trade-off. Top: increasing volatility decreases the marginal payoff of learning (from lighter to darker gray curves). Bottom: increasing volatility decreases the optimal learning precision that maximizes the difference between payoff and cost (from lighter to darker blue curves).

non-greedy decisions are triggered by random noise in the learning process, rather than by an overt drive to seek information during choice.

This finding requires a reconsideration of the very nature of non-greedy decisions in volatile environments. In addition to noise in action selection and lapses in attention, which were both negligible in our task (see Supplementary Modeling), these non-greedy decisions have traditionally been regarded as exploratory and information-seeking, and they should therefore happen only when there is uncertainty regarding the current value of recently unchosen actions. In accordance with this view, we found that the fraction of non-greedy decisions labeled as choice-driven by our noisy RL model depends critically on the absence of knowledge about the outcome of the foregone action on each trial. By contrast, noise-driven variability in action values did not depend on knowledge about the foregone action, suggesting that it reflects a core characteristic of human learning rather than a feature that can be suppressed when the resulting behavioral variability is not useful[14]. In this sense, learning noise resembles the internal corruptive noise found in canonical decision-theoretic models, ranging from signal detection theory to sampling-based theories of inference[3,25]. The moderate consistency of human decisions across identical blocks excluded the possibility that learning noise is primarily due to a misspecification of our learning model[6,26].

The analysis of BOLD signals provided further information about the neural mechanisms underlying the observed learning noise. BOLD activity in a well-documented cognitive control network centered on the dACC correlates positively with trial-to-trial deviations from the exact application of the canonical Rescorla–Wagner rule, even when participants were cued to select a randomly determined action, and thus did not choose. These neural correlates of learning noise differ fundamentally from neural correlates of computational quantities associated with RL (for example, prediction errors or expected values), in the sense that learning noise is neither computed explicitly by our noisy RL model nor thought to be represented in any brain region. Our noisy RL model updates action values with a limited precision, and thus trial-to-trial deviations of each update step from the average reflect the effective variance (inverse precision) of the learning rule. Therefore, neural correlates of learning noise should be interpreted as brain regions with activations that scale with the effective variability of learning steps, not as brain regions that encode or represent learning noise explicitly.

Several previous studies have linked the frontal cortex to exploration and foraging across species[3,11,19], but the specific contributions of different regions have remained unclear. Although several regions showed larger responses during switches, our brain–behavior analysis revealed that only dACC fluctuations exhibit psychometric signatures of learning noise. Such a positive relationship between dACC activity and learning noise may seem at first inconsistent with a causal role of this region in learning[27,28]. However, several recent findings support the idea of learning-specific variability triggered by the dACC. At the theoretical level, the metaplastic synapses in the dACC that are thought to account for adaptive learning in volatile environments go through stochastic transitions between states of faster and slower learning[29]. Neural circuits endowed with such synaptic properties should reflect prediction errors at multiple time scales, as recently observed in the dACC[30], and produce behavioral variability with the same statistical signatures as learning noise.

An intriguing possibility is that computational noise may confer beneficial properties to learning in volatile[5] or high-dimensional environments, as shown in the machine learning literature[31,32]. The structure of learning noise has choice-stabilizing properties and produces an adjustment of exploration to surprise without requiring its explicit monitoring. Beyond these intrinsic benefits, computational noise in RL may also optimize a second trade-off between the marginal payoff of a computation and the cost associated with performing the computation at a certain precision. The dACC has recently been proposed to reflect a similar process by monitoring an expected value of control, defined as the difference between expected payoff and associated cost (cognitive conflict, in particular)[18]. Instead of assuming that patterns of dACC activity explicitly represent such cost, we propose that the cost associated with a computation may be reflected implicitly by its precision (Fig. 8a). This proposal provides a natural explanation as to why learning is subject to limited computational precision, but also makes testable predictions. In particular, increasing volatility reduces the marginal payoff of learning (which, in the limit case, tends towards zero) and thus decreases the precision, which optimizes the payoff–cost trade-off (Fig. 8b). We therefore predict that participants should feature not only increased learning rates[33], but also increased learning noise, in more volatile environments. In this view, the increased dACC activity observed at increased levels of volatility[31] would be a signature of the increased learning noise predicted in such conditions.

Based on previous findings, we reasoned that learning noise may be linked to the ongoing state of the LC–NE system[20,34]. Indeed, LC neurons receive strong projections from the dACC, which in turn produce gain control in several frontal regions implicated in reward-guided learning[24,35]. Pupil-linked fluctuations of the LC–NE state are associated with task disengagement and non-greedy decisions[36,37]. Existing theories have interpreted these findings as evidence of a role for the LC–NE system in controlling the exploration–exploitation trade-off, for which there is only partially conclusive evidence so far[23,36]. We proposed that the LC–NE system may instead mediate the relationship between dACC activity and learning noise. In line with this hypothesis, we observed that

pupillary dilation predicts not only the value of switching away from the previous action, but also the sensitivity of participants to action values. This relationship between pupillary dilation and sensitivity supports the idea that the LC–NE system controls the proposed payoff–cost trade-off by adjusting the computational precision of learning. The recent observation that pupillary dilation increases at high levels of volatility[38] (that is, when the optimal learning precision decreases) provides indirect evidence for this idea, which should be tested formally in future work.

Together, our findings reveal a large source of behavioral variability in reward-guided decision-making, driven by computational noise in the underlying learning process. This noise-driven source of non-greedy decisions is independent of the arbitration between the exploitation of better-valued actions against the exploration of more uncertain ones. As we have shown, the decomposition of non-greedy decisions into noise- and choice-driven components has important consequences for understanding the mechanisms underlying reward-guided behavior and its neurophysiological substrates. Existing models of learning should be revised to allow for noise in their core computations, and include a cost–benefit trade-off regulating their precision.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-019-0518-9.

Received: 11 October 2018; Accepted: 17 September 2019;
Published online: 28 October 2019

References

1. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, 1998).
2. Rescorla, R. A. & Wagner, A. R. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. in *Classical Conditioning II* (eds Black, A. H.Prokasy, W. F.) 64–99 (Appleton-Century-Crofts, 1972).
3. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
4. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore–exploit dilemma. *J. Exp. Psychol. Gen.* **143**, 2074–2081 (2014).
5. Wyart, V. & Koechlin, E. Choice variability and suboptimality in uncertain environments. *Curr. Opin. Behav. Sci.* **11**, 109–115 (2016).
6. Drugowitsch, J., Wyart, V., Devauchelle, A.-D. & Koechlin, E. Computational precision of mental inference as critical source of human choice suboptimality. *Neuron* **92**, 1398–1411 (2016).
7. Fechner, G. T. *Elements of Psychophysics* (Holt, Reinehart & Winston, 1966).
8. Churchland, M. M. et al. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nat. Neurosci.* **13**, 369–378 (2010).
9. Johnson, K. O., Hsiao, S. S. & Yoshioka, T. Neural coding and the basic law of psychophysics. *Neuroscientist* **8**, 111–121 (2002).
10. Palminteri, S., Wyart, V. & Koechlin, E. The importance of falsification in computational cognitive modeling. *Trends Cogn. Sci.* **21**, 425–433 (2017).
11. Boorman, E. D., Behrens, T. E. J., Woolrich, M. W. & Rushworth, M. F. S. How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron* **62**, 733–743 (2009).
12. Palminteri, S., Khamassi, M., Joffily, M. & Coricelli, G. Contextual modulation of value signals in reward and punishment learning. *Nat. Commun.* **6**, 8096 (2015).
13. Lau, B. & Glimcher, P. W. Dynamic response-by-response models of matching behavior in rhesus monkeys. *J. Exp. Anal. Behav.* **84**, 555–579 (2005).
14. Gershman, S. J., Pesaran, B. & Daw, N. D. Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *J. Neurosci.* **29**, 13524–13531 (2009).
15. Yu, A. J. & Cohen, J. D. Sequential effects: superstition or rational behavior? *Adv. Neural Inf. Process. Syst.* **21**, 1873–1880 (2009).
16. Cohen, J. D., McClure, S. M. & Yu, A. J. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **362**, 933–942 (2007).
17. Doya, K. Modulators of decision making. *Nat. Neurosci.* **11**, 410–416 (2008).
18. Shenhav, A., Botvinick, M. M. & Cohen, J. D. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron* **79**, 217–240 (2013).
19. Donoso, M., Collins, A. G. E. & Koechlin, E. Foundations of human reasoning in the prefrontal cortex. *Science* **344**, 1481–1486 (2014).
20. Aston-Jones, G. & Cohen, J. D. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.* **28**, 403–450 (2005).
21. Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J. & Aston-Jones, G. The role of locus coeruleus in the regulation of cognitive performance. *Science* **283**, 549–554 (1999).
22. Eldar, E., Cohen, J. D. & Niv, Y. The effects of neural gain on attention and learning. *Nat. Neurosci.* **16**, 1146–1153 (2013).
23. Jepma, M. & Nieuwenhuis, S. Pupil diameter predicts changes in the exploration-exploitation trade-off: evidence for the adaptive gain theory. *J. Cogn. Neurosci.* **23**, 1587–1596 (2011).
24. Joshi, S., Li, Y., Kalwani, R. M. & Gold, J. I. Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron* **89**, 221–234 (2015).
25. Gershman, S. J. A unifying probabilistic view of associative learning. *PLoS Comput. Biol.* **11**, e1004567 (2015).
26. Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E. & Pouget, A. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron* **74**, 30–39 (2012).
27. Kennerley, S. W., Walton, M. E., Behrens, T. E. J., Buckley, M. J. & Rushworth, M. F. S. Optimal decision making and the anterior cingulate cortex. *Nat. Neurosci.* **9**, 940–947 (2006).
28. Tervo, D. G. R. et al. Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell* **159**, 21–32 (2014).
29. Farashahi, S. et al. Metaplasticity as a neural substrate for adaptive learning and choice under uncertainty. *Neuron* **94**, 401–414.e6 (2017).
30. Meder, D. et al. Simultaneous representation of a spectrum of dynamically changing value estimates during decision making. *Nat. Commun.* **8**, 1942 (2017).
31. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
32. Bottou, L. Large-scale machine learning with stochastic gradient descent. in *Proceedings of COMPSTAT'2010* (eds Lechevallier Y. & Saporta G.) 177–186 (2010).
33. Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).
34. Yu, A. J. & Dayan, P. Uncertainty, neuromodulation, and attention. *Neuron* **46**, 681–692 (2005).
35. Arnsten, A. F. T. & Goldman-Rakic, P. S. Selective prefrontal cortical projections to the region of the locus coeruleus and raphe nuclei in the rhesus monkey. *Brain Res.* **306**, 9–18 (1984).
36. Warren, C. M. et al. The effect of atomoxetine on random and directed exploration in humans. *PLoS One* **12**, e0176034 (2017).
37. Kane, G. A. et al. Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task. *Cogn. Affect. Behav. Neurosci.* **17**, 1073–1083 (2017).
38. Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X. & Bishop, S. J. Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nat. Neurosci.* **18**, 590–596 (2015).

Methods

Participants. Experiment 1 (neuroimaging) included 30 participants (16 women, age 26.0 ± 5.5 years, all right-handed). Participants had no history of neurological or psychiatric disease and had normal or corrected-to-normal vision. One participant was excluded from all analyses due to chance-level performance. Experiment 2 (behavioral) included 30 participants (17 women, age 23.6 ± 4.6 years). Six participants were excluded from pupillometric analyses because of a large amount of missing data (eye blinks, head movements). Experiment 3 (behavioral) included 30 participants (19 women, age 24.2 ± 3.7 years). No statistical methods were used to predetermine sample sizes but our sample sizes are similar to, or larger than, those reported in previous publications[3,6,12]. All tested participants gave written informed consent before taking part in the study, which received ethical approval from relevant authorities (Comité de Protection des Personnes Ile-de-France VI, ID RCB: 2007-A01125-48, 2017-A01778-45). Participants received fixed monetary compensation for their participation in the study, and an additional bonus of up to €10 based on the number of points earned at the end of the experiment.

Experimental protocol. In all three experiments, we asked participants to play a restless, two-armed bandit game. In practice, participants were asked to maximize their monetary payoff by sampling repeatedly from one of two reward sources depicted by colored shapes. Experiments were divided into short blocks of trials, each involving a new pair of colored shapes (56 trials per block in experiments 1 and 3, 96 trials per block in experiment 2). In each trial, participants were asked to choose one of the two shapes presented to the left and right of fixation by pressing one of two buttons with their left or right index finger, and then observed its associated outcome (Fig. 1a). Participants were asked to favor precision over speed, and no time limit was imposed on the latency of their responses (for task instructions, see Supplementary Note).

The payoffs that could be obtained from either shape (from 1 to 99 points) were sampled from probability distributions with means that followed a random walk process (Fig. 1b). More precisely, the mean payoff on trial t $\hat{r}_t$ was sampled from a beta distribution with shape parameters $\alpha = 1 + \hat{r}_{t-1}\exp(\tau)$ and $\beta = 1 + (1 - \hat{r}_{t-1})\exp(\tau)$. This parameterization corresponds to a mode equal to $\hat{r}_{t-1}$ and a spread growing monotonically with τ, fixed to 3.0 across all three experiments. The obtained payoff on trial t r_t was sampled from another beta distribution with shape parameters $\alpha = 1 + \hat{r}_t\exp(\omega)$ and $\beta = 1 + (1 - \hat{r}_t)\exp(\omega)$. This parameterization corresponds to a mode equal to $\hat{r}_t$ and a spread growing monotonically with ω, fixed to 1.5 in experiments 1 and 3, and varied between 1.0 and 2.0 (counter-balanced across blocks) in experiment 2. Parameter τ controls the long-term volatility of the random walk process followed by mean payoffs, whereas parameter ω controls the instantaneous uncertainty about the mean payoff $\hat{r}_t$ based only on the observation of the payoff r_t obtained in trial t.

In experiments 1 and 2, for half of the blocks (four of 8), the reward obtained in each trial from the chosen shape was presented simultaneously with the foregone reward that could have been obtained from the unchosen shape—referred to as the complete outcome condition[12]. For the other half, only the obtained reward was presented—referred to as the partial outcome condition. In experiment 3 (16 blocks in total), participants were always presented with both obtained and foregone rewards (that is, the complete outcome condition). Unknown to participants, half of the blocks (replay blocks) corresponded to exact repetitions of reward sequences presented in the other half (seed blocks, Fig. 3a): the fifth block was an exact repetition of the first block, the sixth block an exact repetition of the second block, and so on. Experiment 1 included not only the choice trials described above, in which participants could select the option they wanted to sample, but also cued trials (25% of all trials, Fig. 6a), in which participants were required to select one of the two shapes (pre-selected randomly).

In experiments 1 and 2, partial outcome and complete outcome blocks were presented in a pseudo-randomized order across participants. The trajectories of mean and obtained payoffs were generated independently for each participant and each block using the procedure described above. During data collection, tested participants were not blinded to the experimental condition (that is, they could either clearly observe or not observe the foregone reward that could have been obtained from the unchosen shape), whereas experimenters were blinded (that is, they were not in the room during experimental blocks). None of the analyses was performed blind to the experimental conditions.

Computational model. To characterize the origin of non-greedy decisions made in this task, we derived a RL model in which the Rescorla–Wagner rule applied to update the value (expected reward) Q_{t-1} associated with the chosen action a_{t-1} is corrupted by additive random noise ε_t:

$$Q_t = Q_{t-1} + \alpha(r_{t-1} - Q_{t-1}) + \varepsilon_t$$

where α is the learning rate used to update action values based on the prediction error between obtained reward r_{t-1} and expected reward Q_{t-1} on the previous trial, and ε_t is drawn from a normal distribution with zero mean and s.d. σ_t equal to a constant fraction ζ of the magnitude of the prediction error: $\sigma_t = \zeta|r_{t-1} - Q_{t-1}|$. This noisy learning rule reduces to the exact (noise-free) Rescorla–Wagner rule when $\zeta \to 0$.

As in existing theories, we modeled the choice process using a stochastic softmax action selection policy, controlled by an inverse temperature β and an optional choice hysteresis ξ:

$$a_t \sim B\left(\frac{1}{1 + \exp\left(-\beta\left(Q_{t,A} - Q_{t,B}\right) - \xi\,\mathrm{sign}(a_{t-1})\right)}\right)$$

where B(.) denotes the Bernoulli distribution, and $Q_{t,A}$ and $Q_{t,B}$ correspond to the values associated with actions A and B, coded as +1 for A and −1 for B. This stochastic action selection policy reduces to a purely greedy (value-maximizing) argmax policy when $\beta \to \infty$. For a more detailed description of the computational model, see Supplementary Modeling.

Model fitting procedure. Fits for all models were based on Monte Carlo methods[39]. More precisely, for the exact RL models (without learning noise), we used an iterated batch importance sampler (IBIS)[40]. IBIS is a sequential Monte Carlo (SMC) algorithm for exploring a sequence of parameter posterior distributions when the likelihoods $p(a_t|a_{1:t-1}, r_{1:t-1}, \theta)$, where θ corresponds to model parameters $\{\alpha, \beta, \xi\}$, are tractable. This class of Monte Carlo methods could not be used for the noisy RL models (with non-zero learning noise) because the corresponding likelihoods become intractable in this case. We thus used the SMC^2 algorithm[41] to perform parameter inference for the noisy RL models. Technical details about these model fitting procedures can be found in Supplementary Modeling. Some analyses required further estimation of the smoothing distributions of the trajectories of action values over the course of each block $p(Q_{1:n}|a_{1:n}, r_{1:n}, \theta^{\mathrm{MAP}})$, where n corresponds to the number of trials in each block, and θ^{MAP} to maxima a posteriori for model parameters $\{\alpha, \zeta, \beta, \xi\}$. To obtain samples approximately distributed under the smoothing distributions, we used the forward filter backward simulator (FFBSi)[42,43] to obtain K samples $\tilde{Q}_{1:n,k}$.

Model recovery procedure. We implemented a model recovery procedure to test the robustness of our model fitting and selection procedures. The recovery procedure consists of simulating our three candidate models of interest (model 1, exact RL with softmax policy; model 2, noisy RL with softmax policy; model 3, noisy RL with argmax policy), and applying our model fitting and selection procedures to obtained simulations to test whether we can accurately recover the simulated model. We simulated each model 29 times (once for each subject in experiment 1) and used the posterior means obtained by fitting the models to each participant as parameter values. Thus, for model 1 (exact RL with softmax policy) and 3 (noisy RL with argmax policy), we set parameter values to the posterior means obtained by fitting models 1 and 3 to each subject. For model 2 (noisy RL with softmax policy), the softmax and learning noise parameters were set to best-fitting values obtained with models 1 and 3, respectively. We decided to use these parameter values to produce simulations in which the two sources of variability in the model (that is, learning noise and choice stochasticity) generated the same amount of behavioral variability. This recovery procedure provides an external validation for the tested models: their sources of variability are recoverable from behavior.

Computing the mutual information of successive decisions. We estimated the mutual information MI_t of successive actions $\{a_{t-1}, a_t\}$ as a model-free behavioral metric to distinguish best fits of exact (noise-free) and noisy RL models of human behavior using the following standard equation:

$$MI_t = \sum_{a_{t-1} \in \{A,B\}} \sum_{a_t \in \{A,B\}} p(a_{t-1}, a_t) \log\left(\frac{p(a_{t-1}, a_t)}{p(a_{t-1})p(a_t)}\right)$$

For both exact and noisy RL models, we simulated decisions using parameter values fitted to the behavioral data, and compared the trade-off between the fraction of non-greedy decisions and the mutual information between consecutive decisions. The exact and noisy RL models predicted quantitatively different trade-offs between these two metrics, to falsify the exact RL model, which could not account for the trade-off obtained using human data.

Distinguishing learning noise from choice stochasticity. In experiments 1 and 2, we estimated the fraction of non-greedy decisions that could be accounted for by learning noise. First, to label a decision as non-greedy, we fitted the exact RL model to the behavior of the participants and labeled every decision for which action values $\{Q_{t,A}, Q_{t,B}\}$ predicted by the exact model favored the unchosen action as non-greedy[3]. Second, we fitted the noisy RL model to the behavior of participants and estimated the smoothing distributions of action values throughout each block. We then labeled as noise-driven the non-greedy decisions for which noisy realizations of the learning rule resulted in an opposite ranking of action values $\left\{\tilde{Q}_{t,A}, \tilde{Q}_{t,B}\right\}$ to exact (noise-free) applications of the same rule:

$$\mathrm{sign}\left(Q_{t,A} - Q_{t,B}\right) \neq \mathrm{sign}\left(\tilde{Q}_{t,A} - \tilde{Q}_{t,B}\right)$$

We report in the main text the fraction of non-greedy decisions labeled as noise-driven, which would approach 1 for a noisy learner relying on a deterministic argmax action selection policy, and 0 for an exact learner relying on a stochastic softmax action selection policy.

For the estimation of the behavioral variability generated by learning noise and choice stochasticity separately, we used the following procedure. First, we estimated the behavioral variability generated by learning noise alone by computing the s.d. of the difference between noisy action values $\tilde{Q}_t$ obtained through smoothing distributions and action values $\hat{Q}_t$ obtained through exact (noise-free) applications of the Rescorla–Wagner rule to noisy action values in the previous trial $\tilde{Q}_{t-1}$. Second, we estimated the behavioral variability generated by choice stochasticity alone by approximating the logistic (softmax) distribution as the cumulative probability density function of a normal distribution (see Supplementary Modeling for the full analytical derivation). From this approximation, we obtained the s.d. that best approximates a logistic distribution of inverse temperature β as $\pi/(\beta\sqrt{3})$.

Obtaining the bias–variance decomposition of learning noise. To decompose learning noise into systematic and random deviations from the Rescorla–Wagner rule, we estimated the consistency of human decisions across repetitions of the exact same sequence of rewards (experiment 3). We restricted this experiment to the complete outcome condition for two important reasons. First, we wanted behavioral variability to be driven solely by learning noise and not by a softmax action selection policy (noisy versus exact RL, BF $\approx 10^{411.8}$, exceedance $P > 0.999$; argmax versus softmax policy, BF $\approx 10^{6.3}$, exceedance $P = 0.997$). Second, we wanted the action values predicted by an exact RL model to be identical across the two repetitions of the same block, irrespective of the decisions made in the two repeated blocks. This was ensured by the observation that, as in previous experiments, participants learned equally from obtained (chosen) and foregone (unchosen) rewards in this additional dataset (learning rate α, chosen, 0.571 ± 0.036; unchosen, 0.597 ± 0.038, paired t-test, $t_{29} = 1.6$, $P = 0.123$, $BF_{H0} = 1.7$). Beyond this absence of difference in mean learning rates, validated by BMS (BF $\approx 10^{16.0}$, exceedance $P > 0.999$), learning rates associated with chosen and unchosen actions were also highly correlated across participants (linear correlation, $r^2 = 0.821$, d.f. $= 28$, $P < 0.001$). This pattern of findings indicates that the action values predicted by exact RL in a given trial do not depend on decisions made in earlier trials (which are likely to differ to some extent across the two repetitions of the same block).

We then relied on the observed consistency of human decisions across repeated blocks to split learning noise into a predictable bias term reflecting systematic deviations from the Rescorla–Wagner rule and an unpredictable variance term, reflecting random deviations around this canonical rule. Indeed, systematic deviations tend to increase the consistency of decisions across repeated blocks, whereas random deviations tend to decrease this consistency. In practice, we fitted the noisy RL model to each participant and then simulated versions of the model in which learning noise was split into two additive terms: a predictable bias term with noise realizations that were duplicated in the two repetitions of each block, and an unpredictable variance term with noise realizations that were sampled independently across repeated blocks. We varied this bias–variance trade-off from zero (fully predictable) to one (fully unpredictable) in 100 equally spaced steps for the simulations of each participant, and found the split that best accounted for the consistency of human decisions across repeated blocks.

Computing choice sensitivity to action values. We estimated the choice sensitivity of participants to action values by fitting a logistic regression model to the decision of each participant in trial t $a_t \in \{A,B\}$ as a function of the difference in action values $\hat{Q}_{t,A} - \hat{Q}_{t,B}$ predicted by exact applications of the Rescorla–Wagner rule to noisy action values $\tilde{Q}_{t-1,A} - \tilde{Q}_{t-1,B}$ in the previous trial $t-1$:

$$p(a_t = A) = \Phi\left(\beta_0 + \beta_1\left(\hat{Q}_{t,A} - \hat{Q}_{t,B}\right)\right)$$

where β_0 corresponds to a choice bias, β_1 corresponds to the choice sensitivity to action values and $\Phi(\cdot)$ corresponds to the sigmoid function. We used the action values $\hat{Q}_{t,A}$ and $\hat{Q}_{t,B}$ predicted by exact (noise-free) applications of the Rescorla–Wagner rule on the last update step, such that the decision variability in trial t was unaccounted for by action values $\hat{Q}_{t,A}$ and $\hat{Q}_{t,B}$, and estimated by the choice sensitivity β_1 to action values.

Computing regressors for model-based neuroimaging and pupillometry. In experiment 1, we regressed canonical choice- and outcome-locked BOLD responses at each voxel against four trial-wise quantities derived from the noisy RL model fitted to the behavior of each participant: the similarity between action values following each update step (choice conflict), the difference between chosen and unchosen action values (choice value), the prediction error associated with the obtained reward, and the magnitude of learning noise corrupting the associated update of action values. This final quantity, which was specific to our noisy RL model, was computed as the predicted deviation $|\varepsilon_t|$ of noisy action values following each update step from the exact application of the Rescorla–Wagner rule to the same update step:

$$|\varepsilon_t| = \left|\tilde{Q}_t - \left(\tilde{Q}_{t-1} + \alpha \cdot \left(r_{t-1} - \tilde{Q}_{t-1}\right)\right)\right|$$

where $\tilde{Q}_t$ refers to noise-corrupted action values predicted by the noisy RL model conditioned on all observed rewards $r_{1:n}$ and all actions $a_{1:n}$ made by the participant. In practice, because summary statistics for $\tilde{Q}_t$ cannot be derived analytically, particle smoothing procedures were used to draw samples from the posterior distributions and $|\varepsilon_t|$ was averaged across drawn samples (see Supplementary Modeling for more details). Importantly, the corresponding general linear model was constructed using sequential orthogonalization to ensure that the noise regressor captured residual BOLD variance unaccounted for by the previous regressors, which were also predicted by the exact RL model. In experiment 2, we regressed phasic pupillary dilation at each time sample around the onset of each choice period against three trial-wise quantities derived from the noisy RL model fitted to the behavior of each participant: choice conflict, choice value and the magnitude of learning noise corrupting the preceding update of action values. This final quantity was defined as described above for the model-based analysis of BOLD signals. We omitted the prediction error associated with the obtained reward in pupillometric analyses because it was not expected to trigger reliable modulations of pupillary dilation.

Neuroimaging data acquisition and preprocessing. A Siemens Prisma 3 T scanner (Centre de Neuroimagerie de Recherche, Paris, France) and a 64-channel head coil were used to acquire both high-resolution T1-weighted anatomical MRI using a 3D MPRAGE with a resolution of 1 mm^3 (isometric) and T2*-weighted multiband-echo planar imaging (mb-EPI) with a multiband factor of 3 and an acceleration factor of 2 (GRAPPA). Parameters for fMRI time-series acquisition were as follows: 54 slices acquired in ascending order, an isometric voxel size of 2.5 mm, a repetition time of 1.1 s and an echo time of 25 ms. A tilted plane acquisition sequence was used to optimize sensitivity to BOLD signal in the orbitofrontal cortex[44,45]. Preprocessing included co-registration of the anatomical T1 images with the mean EPI, segmentation and normalization to a standard T1 template, and averaging across participants to allow group-level anatomical localization.

Preprocessing of functional mb-EPI sequences consisted of spatial realignment, movement correction, reconstruction and distortion correction and normalization using the same transformation applied to structural T1 images. Normalized images were spatially smoothed using a Gaussian kernel with a full width at a half maximum of 8 mm. All preprocessing steps except distortion correction were performed using SPM12 (Wellcome Trust Center for Neuroimaging, London, UK; www.fil.ion.ucl.ac.uk). Distortion correction was performed using image unwarping and reconstruction as implemented in the FMRIB Software Library (FSL)[46].

Predicting behavioral variability from neurophysiological signals. We formulated a brain–behavior analysis to predict behavioral variability on the basis of neurophysiological signals: trial-to-trial BOLD fluctuations in five ROIs (the dACC, the right dlPFC, the PPC, the FPC and the vmPFC) in experiment 1, and trial-to-trial pupillary fluctuations in experiment 2. In both experiments, we first standardized the decision variable used by each participant (corresponding to a decision sensitivity of one and an unbiased decision criterion) by fitting a standard logistic regression model to the decision of each participant in trial t $a_t \in \{A,B\}$ as a function of the difference in action values $\hat{Q}_{t,A} - \hat{Q}_{t,B}$ predicted by exact applications of the Rescorla–Wagner rule to noisy action values $\tilde{Q}_{t-1,A} - \tilde{Q}_{t-1,B}$ in the previous trial $t-1$:

$$p(a_t = A) = \Phi\left(\beta_0 + \beta_1\left(\hat{Q}_{t,A} - \hat{Q}_{t,B}\right)\right)$$

where β_0 and β_1 correspond to the two fitted parameters of the logistic regression model. Note that we used the action values predicted by exact (noise-free) applications of the Rescorla–Wagner rule on the last update step, such that neurophysiological signals could be used to predict the behavioral effect of learning noise on the last update step. We could then compute a standardized decision variable corresponding to the adjusted difference in action values $\Delta Q^* = \beta_0^* + \beta_1^*\left(\hat{Q}_{t,A} - \hat{Q}_{t,B}\right)$, where β_0^* and β_1^* correspond to best-fitting parameter values. This first standardization step allowed the average sensitivity of each participant to action values to be set to the same value. This adjusted difference in action values ΔQ^* was then used in all brain–behavior analyses described below.

In experiment 1, we used a logistic regression model to predict the decision of each participant to select either action as a function of the adjusted difference in action values ΔQ^*, and single-trial deconvolved BOLD responses in the five ROIs. The use of such a forward model of behavior, in which BOLD responses in the different ROIs can be included simultaneously, enables their shared variance to be accounted for. We entered single-trial BOLD responses x_t in the model as modulators of the sensitivity of participants to the adjusted difference between action values in the form of an interaction term $x_t * \Delta Q^*$. We followed a fully factorial scheme by constructing and estimating the posterior probabilities associated with all possible combinations of ROIs ($32 = 2^5$) using maximum likelihood estimation. We could then estimate the family-wise probability of each ROI to modulate sensitivity independently of the involvement of other ROIs. Note that, in contrast to posterior probabilities, these family-wise probabilities do not sum to one: they would all equal zero if no ROI modulates sensitivity, and all equal one if all ROIs modulate sensitivity.

In experiments 1 and 3, we further tested whether the negative modulation of sensitivity by dACC responses and pupillary fluctuations, respectively, could be caused not by random fluctuations in action values (as predicted by learning noise), but rather by directed fluctuations in the value of switching away from the previous action (as predicted by adjustments of the exploration–exploitation trade-off). To test this alternative brain–behavior relationship, we modified the logistic regression model to predict the decisions of participants to switch away from their previous action (that is, $a_t \neq a_{t-1}$) as a function of the difference between the value of switching $\hat{Q}_{t,\text{switch}} = \hat{Q}_t(a_t \neq a_{t-1})$ and the value of repeating the previous action $\hat{Q}_{t,\text{repeat}} = \hat{Q}_t(a_t = a_{t-1})$. We used the same standardization step described above to compute the adjusted difference in action values ΔQ^*, this time between switching and repeating the previous action rather than between actions A and B. We then entered single-trial neurophysiological signals (either dACC responses or pupillary dilation) at trial t x_t in the model not only as a modulator of the sensitivity of participants to the value difference between switching and repeating in the form of an interaction term $x_t{}^*\Delta Q^*$, but also as a modulator of the value of switching in the form of an additive term x_t. Because the same neurophysiological signal could simultaneously predict random and directed effects, we again followed a fully factorial scheme by constructing and estimating the model evidence associated with the $4 = 2^2$ combinations of the two possible modulations using maximum likelihood estimation.

Statistical procedures. Standard paired *t*-tests were used to compare conditions at the group level. Data distribution (individual data points shown on the main figures) was assumed to be approximately normal, and was therefore not formally tested for normality. When standard *t*-tests yielded non-significant results with a meaningful interpretation, we performed an additional Bayesian test to obtain the BF_{H0} throughout the main text. This Bayesian test was performed using the BayesFactor library in the R language, using default priors[47].

BMS analyses used the unbiased estimate of the marginal likelihood obtained from the model fitting procedure as the model evidence metric (see Supplementary Modeling). This metric integrates over parameters and thus penalizes model complexity without requiring an explicit penalization term. BMS was conducted using separate fixed-effects and random-effects approaches. The fixed-effects approach assumes that all participants are relying on the same model, and consists of comparing the log-marginal likelihood summed across participants for each tested model. By contrast, the random-effects approach assumes that different participants may rely on different models, and consists of estimating the Dirichlet distribution over models from which participants draw[48], as implemented in the SPM12 software package (Statistical Parametric Mapping).

Reporting Summary. Further information on the research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data (behavioral, neuroimaging and pupillometric) that support these findings are available from the corresponding author upon request.

Code availability

Python and C++ code for fitting all computational models described in the article are available at https://github.com/csmfindling/learning_variability. The algorithmic backbone of the Monte Carlo procedures used to fit models can be found in Supplementary Modeling Note.

References

39. Robert, C. & Casella, G. *Monte Carlo Statistical Methods* (Springer, 2004).
40. Chopin, N. A sequential particle filter method for static models. *Biometrika* **89**, 539–552 (2002).
41. Chopin, N., Jacob, P. E. & Papaspiliopoulos, O. SMC^2: an efficient algorithm for sequential analysis of state space models. *J. R. Stat. Soc. B* **75**, 397–426 (2013).
42. Lindsten, F. & Schön, T. B. Backward simulation methods for Monte Carlo statistical inference. *Found. Trends Mach. Learn.* **6**, 1–143 (2013).
43. Doucet, A., Godsill, S. & Andrieu, C. On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. Comput.* **10**, 197–208 (2000).
44. Deichmann, R., Gottfried, J., Hutton, C. & Turner, R. Optimized EPI for fMRI studies of the orbitofrontal cortex. *Neuroimage* **19**, 430–441 (2003).
45. Weiskopf, N., Hutton, C., Josephs, O. & Deichmann, R. Optimal EPI parameters for reduction of susceptibility-induced BOLD sensitivity losses: a whole-brain analysis at 3T and 1.5T. *Neuroimage* **33**, 493–504 (2006).
46. Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W. & Smith, S. M. FSL. *Neuroimage* **62**, 782–790 (2012).
47. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* **16**, 225–237 (2009).
48. Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *NeuroImage* **15**, 1004–1017 (2009).

Acknowledgements

We thank C. Summerfield (University of Oxford; Google DeepMind) for comments on an earlier version of the manuscript. This work was supported by a starting grant from the European Research Council awarded to V.W. (ERC-StG-759341), a junior researcher grant from the Agence Nationale de la Recherche awarded to V.W. (ANR-14-CE13-0028) and two department-wide grants from the Agence Nationale de la Recherche (ANR-10-LABX-0087 and ANR-10-IDEX-0001-02 PSL). C.F. was supported by a graduate research fellowship from the Direction Générale de l'Armement (2015-60-0041). S.P. was supported by a CNRS-Inserm ATIP-Avenir grant (R16069JS) and a research grant from the Programme Emergence(s) of the City of Paris.

Author contributions

S.P. and V.W. were responsible for conceptualization. C.F., V.W. and S.P. were responsible for the methodology. C.F., V.S. and V.W. performed the formal analysis. V.S. and R.D. carried out the investigations. C.F., V.S. and V.W. wrote the original draft. C.F., V.S., S.P. and V.W. reviewed and edited the report. V.W. supervised the study and acquired funding.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-019-0518-9.

Correspondence and requests for materials should be addressed to V.W.

Peer review information *Nature Neuroscience* thanks Samuel Gershman, Yonatan Loewenstein, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Corresponding author(s): Valentin Wyart
valentin.wyart@ens.fr

Last updated by author(s): Sep 4, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	MATLAB R2017b/Psychtoolbox-3 (behavior), EYELINK II CL v4.594 (pupillometry)
Data analysis	FSL 6.0, SPM12, MATLAB R2017b, Python 2.7, C++11

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support these findings are available from authors upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The experimental paradigm used in this study is a restless, two-armed bandit task. The behavioral data collected were quantitative and consisted of participants' choices and reaction times. Additionally, we collected quantitative whole-brain BOLD signals (in experiment 1) and pupil dilation (in experiment 3).
Research sample	Experiment 1 (n = 30): 16 females, age: 26.0 +/– 5.5 years (behavior + fMRI) Experiment 2 (n = 30): 17 females, age: 23.6 +/– 4.6 years (behavior + pupillometry) Experiment 3 (n = 30): 19 females, age: 24.2 +/– 3.7 years (behavior) All participants were recruited through the web-based recruitment platform of our university (RISC - http://expesciences.risc.cnrs.fr/). No statistical methods were used to pre-determine sample sizes but our sample sizes are similar or larger to those reported in previous publications (Daw, N. D et al., 2006; Drugowitsch, J., Wyart, V et al., 2016; Palminteri, S., et al. 2015)
Sampling strategy	No power analysis could be performed for this study since the new reinforcement learning model we developed to fit participants' choices had never been tested before. The chosen sample size of n = 30 in all three experiments is matches the commonly accepted good practices in this field. In particular, the chosen sample size was determined a priori for all three experiments.
Data collection	The behavioral data were collected using the Psychtoolbox-3 toolbox for MATLAB; pupillometric signals were collected using an EyeLink eye-tracker; BOLD fMRI data were collected using a Siemens Prisma 3T scanner.
Timing	Experiment 1: Jan 2017 - Mar 2017. Experiment 2: Jan 2016 - Feb 2016. Experiment 3: Nov 2017
Data exclusions	Experiment 1 (behavior & fMRI): 1 participant was excluded because he/she failed to understand task instructions and performed at chance level. Experiment 2 (behavior): no participant was excluded. Experiment 2 (pupillometry) : 6 participants were excluded because low-quality pupillometric signals. Experiment 3 (behavior): no participant was excluded.
Non-participation	No participant cancelled his/her participation from any of the three experiments.
Randomization	Participants were not allocated into distinct experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Experiment 1 (fMRI) : 16F/14M, age: 26.0 +/– 5.5 years, all healthy with no history of neurological or psychiatric disorders. All subjects were right-handed and had normal or corrected-to-normal vision. Experiment 2: 17F/13M, age: 23.6 +/– 4.6 years, all healthy with no history of neurological or psychiatric disorders. All subjects had normal or corrected-to-normal vision. Experiment 3 : 19F/11M, age: 24.2 +/– 3.7 years, all healthy with no history of neurological or psychiatric disorders. All subjects had normal or corrected-to-normal vision.
Recruitment	All participants were recruited through the web-based recruitment platform of our university.
Ethics oversight	Comité de Protection des Personnes Ile-de-France VI, ID RCB: 2007-A01125-48, 2017-A01778-45

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Magnetic resonance imaging

Experimental design

Design type	Experiment 1: task, event-related design for fMRI analyses (see below) Experiment 2: task, event-related design for pupillometric analyses (see below) Experiment 3: task (see below)
Design specifications	Experiment 1: 8 blocks of 56 trials each, partial/complete outcome x free/cued trials Experiment 2: 8 blocks of 96 trials each, partial/complete outcome Experiment 3: 16 blocks of 56 trials each, seed/replay block
Behavioral performance measures	Fraction of trials where the action associated with the largest underlying mean reward is selected.

Acquisition

Imaging type(s)	Functional and structural images
Field strength	3T
Sequence & imaging parameters	High resolution T1-weighted anatomical MRI using a 3D MPRAGE with a resolution of 1mm3 voxel and T2*-weighted multiband-echo planar imaging (mb-EPI) with multi-band factor of 3 and acceleration factor of 2 (GRAPPA).
Area of acquisition	Whole-brain acquisition

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	Preprocessing of the mb-EPI consisted of spatial realignment, movement correction, reconstruction and distortion correction, and normalization using the same transformation as applied for the structural images. Normalized images were spatially smoothed using a Gaussian kernel with a full width at a half-maximum of 8 mm. All the preprocessing except for the distortion correction was done using the SPM12 (Wellcome Trust Center for NeuroImaging, London, UK; ww.fil.ion.ucl.ac.uk). Distortion correction consisted of image unwarping and reconstruction done using FSL software.
Normalization	Normalized, non-linear
Normalization template	MNI305
Noise and artifact removal	Six motion parameters were included in every GLM specified for fMRI BOLD signal analysis to correct for motion artifacts.
Volume censoring	No volumes/scans were excluded from the analyses reported in the article.

Statistical modeling & inference

Model type and settings	First-level : univariate. Second-level : random effects (unless noted otherwise).
Effect(s) tested	One-sample t-tests against zero.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☒ Both

Anatomical location(s): Automatic labeling and probabilistic atlas

Statistic type for inference (See Eklund et al. 2016)	Cluster-wise
Correction	FWE, cluster-wise at whole-brain level

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☒	☐ Multivariate modeling or predictive analysis