Reviewed Preprint
v1 • October 2, 2025
Not revised

Reviewed Preprint
v2 • June 12, 2026
Revised by authors

For correspondence:
Jonas.Simoens@hotmail.com
Senne.Braem@UGent.be
tom.verguts@ugent.be

Competing interests: No competing interests declared

Funding: See page 25

Reviewing editor: Andreea Oliviana Diaconescu, University of Toronto, Canada

Two time scales of adaptation in human learning rates

Jonas Simoens[1], Senne Braem[1], Pieter Verbeke[1], Haopeng Chen[1], Stefania Mattioni[1], Mengqiao Chai[1], Nicolas W Schuck[2], Tom Verguts[1]

[1]Department of Experimental Psychology, Ghent University, Ghent, Belgium • [2]Institute of Psychology, Universität Hamburg, Hamburg, Germany

eLife Assessment

This study makes a **valuable** contribution to the understanding of meta-learning and its neural mechanisms by distinguishing two timescales of learning rate adaptation: rapid, within-block reductions and slower, location-specific, meta-learned adjustments. Behavioural data and computational modelling provide **convincing** evidence that individuals adjust learning rates both rapidly in response to uncertainty and more gradually through meta-learning of environmental statistics. Neuroimaging results indicate that meta-learned learning rates are represented in orbitofrontal cortex, and that prediction errors are encoded across a distributed network including the ventral striatum, where they are modulated by expectations about error magnitude. The manuscript is timely and clearly written and opens the door to future work on how these signals contribute to adaptive behaviour.

https://doi.org/10.7554/eLife.108223.2.sa3

Abstract

Different situations may require radically different information updating speeds (i.e., learning rates). Some demand fast learning rates, while others benefit from using slower ones. To adjust learning rates, decision makers could rely on either global, meta-learned differences between environments, or faster but transient adaptations to locally experienced prediction errors. Here, we introduce a new paradigm that allows researchers to measure and empirically disentangle both forms of adaptations. Participants performed short blocks of trials of a continuous estimation task – fishing for crabs – on six different islands that required different optimal (initial) learning rates. Across two experiments, participants showed fast adaptations in learning rate within a block. Critically, participants also learned global environment-specific learning rates over the time course of the experiment, as evidenced by computational modelling and by the learning rates calculated on the very first trial when revisiting an environment (i.e., unconfounded by transient adaptations). Using representational similarity analyses of fMRI data, we found that differences in voxel pattern responses in the central orbitofrontal cortex correlated with differences in these global environment-specific learning rates. Our findings show that humans adapt learning rates at both slow and fast time scales, and that the central orbitofrontal cortex may support meta-learning by representing environment-specific task-relevant features such as learning rates.

Introduction

Decisions that humans make on a daily basis range from the ordinary, such as what to have for lunch, to the life-defining, such as what career to pursue. The computational framework of reinforcement learning (RL) stipulates that such decision making requires estimating and continuously updating relevant feature values of specific options (e.g., the nutritional value, the

tastiness, and the price of the lunch), and subsequently using these estimates to make a good choice. The most common RL approach to learning such values is the delta rule (Rescorla & Wagner, 1972):

$$Q^a_{t+1} = Q^a_t + \alpha\,(r_t - Q^a_t)$$

where the (Q) value of action a at time $t + 1$ is updated in proportion to the error made in predicting the outcome r of action a at time t (i.e., the prediction error). This learning algorithm requires choosing a learning rate α. Performance of the algorithm depends on the value of this parameter, and different values are optimal for different environments (Sutton & Barto, 2018; Verbeke & Verguts, 2024). Learning the parameters that shape learning is termed meta-learning (Binz et al., 2024; Wang et al., 2018), and theories of human meta-learning suggest that people can flexibly adapt their learning rates to the environment (Mathys, 2011; Schweighofer & Doya, 2003; Silvetti et al., 2018). Similarly, in artificial agents, setting a learning rate is critical, and a breakthrough in AI was the development of an algorithm that set its learning rate in an adaptive manner (Adam optimizer (Kingma & Ba, 2017)).

Consistent with these ideas, humans can adjust their learning rate to the reward volatility and variability of a given environment (Behrens et al., 2007; Browning et al., 2015; Cook et al., 2019; Goris et al., 2021). An important limiting aspect of these studies, however, is that participants typically spend extended periods of time within one environment. Therefore, previous studies cannot dissociate between fast, transient adaptations (i.e., fast time scale learning within an environment; (Bai et al., 2014; Krugel et al., 2009; Nassar et al., 2012) versus learned environment-specific learning rates that can be reused when revisiting an environment (i.e., slow time scale learning about an environment (Simoens et al., 2024)). From an optimality perspective (Kalman filter; (Dayan et al., 2000)), the learning rate should gradually decrease as the task statistics within an environment become increasingly known. When the environmental statistics are reset (e.g., when visiting a new environment), learning rate should be reset as well. Nevertheless, if the higher-order statistics in an environment remain fixed (e.g., amount of noise in the environment), an optimal agent could (meta-)learn higher-order parameters (e.g., the starting point of the learning rate) so that learning in the environment becomes increasingly efficient. From this perspective, people could learn about the optimal learning rate at two levels: On a fast time scale in response to local prediction errors as an environment becomes known, and on a slower time scale as an environment's higher-order statistics and optimal initial settings are learned.

To test this, we administered a novel task in two experiments in which participants went fishing for crabs on six different locations around an island that allowed us to measure both trial-by-trial (transient) adaptations in learning rate as a function of locally experienced prediction errors, as well as learned variations in learning rate adapted to the environmental statistics. This task required continuous responses (i.e., estimating the crab locations) and provided continuous feedback (on those crab locations), which allowed us to estimate learning rates on a trial-by-trial basis by calculating how much people updated their fishing location based on feedback (Nassar et al., 2012). Importantly, each of the six locations around the island had one of three different optimal initial learning rates. This was achieved by changing two quantities that governed the outcome (crab location) distributions of each island: the standard deviation of the distribution that determined the latent mean of crab locations upon an island visit (prior distribution), and the standard deviation of the noise with which crabs were sampled around their latent mean (sampling distribution). As a result, different locations around the island required different learning rates for optimal task performance.

Participants switched between locations on a block-by-block basis, where blocks only lasted two to ten trials (depending on the experiment, see below). Crucially, without prior experience, environments were indistinguishable up to the moment feedback was provided on the second trial of each block, so any differences in learning rates between locations after feedback on the first trial suggest meta-learning of environment-specific learning rates. Across two experiments, we

found that participants dynamically updated their learning rate within environments, but also, critically, learned over time to use different initial learning rates when revisiting different environments, showing the meta-learning of learning rates tailored to the different environments.

In the second experiment, we also collected fMRI data to investigate which brain regions are involved in representing sustained environment-specific learning rates. Namely, we performed representational similarity analyses on neural voxel pattern activity when participants had just been transported to the next location around the island, while they were preparing to perform the task in a particular environment. We hypothesized that the orbitofrontal cortex (OFC) would be the main region involved in this preparation, consistent with its broader role in the representation of task states (Moneta et al., 2024; Schuck et al., 2016; Stalnaker et al., 2015; Wilson et al., 2014)– the integration of contextual information that is necessary to predict the outcomes of decisions, crucial for reward maximization. While shifts in task state representations and value computations are often linked to the OFC, the environment-specific responses to reward predictions are often linked to the basal ganglia. Therefore, we also studied ventral striatum activity as a candidate for processing prediction errors on the first trial of each block. That is, given the role of the ventral striatum in processing (reward) prediction errors (Calderon et al., 2021; O'Doherty et al., 2004; Pessiglione et al., 2006; Schultz et al., 1997), we assumed that this region would show the effect of these learned environment-specific learning rates through a differential response to on-task (reward) prediction errors.

Results

Experiment 1

Fifty participants performed a novel crab-fishing task. At the start of each of 60 blocks, a boat took them to one of six locations around an island (Figure 1A). In each location, participants could drop a cage on a chosen position ten times (trials), with the goal to catch as many crabs as possible (Figure 1C). Each time a cage was dropped, five crabs appeared and spread out evenly from one position in the sand sampled from the sampling distribution $S \sim N\left(\mu_s, \sigma_s^2\right)$, a truncated normal distribution ranging between $\mu_s \pm 1.65 * \sigma_s$. Each crab was either caught by the cage or ran away. At the beginning of each block, the latent mean of the sampling distribution, μ_s, was sampled from the prior distribution $\mu_s \sim N\left(\mu_p, \sigma_p^2\right)$, truncated between $\mu_p \pm 1.65 * \sigma_p$, with μ_p set to be the centre of the screen, where the cage position was initialized on the 1st trial of each block. Crucially, the variance of the latent mean σ_p^2, and the noise variance around that mean σ_S^2, were dependent on the location around the island. In two randomly selected adjacent locations, σ_p^2 was large, while σ_S^2 was small. Here, the true mean position of the crabs was widely dispersed and could be nearly everywhere on the screen, while the individual crabs appeared very close to that true mean. Hence, participants could infer the mean crab position from a single observation and performed best if they strongly adjusted the position of the cage after the first trial. We termed this the *low noise environment*, which required a high initial learning rate (Figure 1B). On the two adjacent locations on the exact opposite side of the island, the situation was reversed (i.e., small σ_p^2, large σ_S^2, henceforth the *high noise environment*), requiring a low initial learning rate. Finally, on the two locations in between, σ_p^2 and σ_S^2, were intermediate and equal (i.e., the *medium noise environment*), requiring an intermediate initial learning rate.

On the first trial of each block, participants could only drop the cage in the centre of the screen. Crucially, the position of the first crabs was determined by the randomly drawn latent mean (with variance σ_p^2), and the variance of the sampling distribution $\left(\sigma_s^2\right)$, i.e. corresponding to a normal distribution with mean equal to the centre of the screen and variance equal to $\sigma_p^2 + \sigma_s^2$. Although σ_p^2 and σ_S^2 differed across the different environments as described above, their sum was constant, and hence the normal distribution for the first trial in a block was identical across the three environments. Therefore, the first prediction errors participants experienced across the three

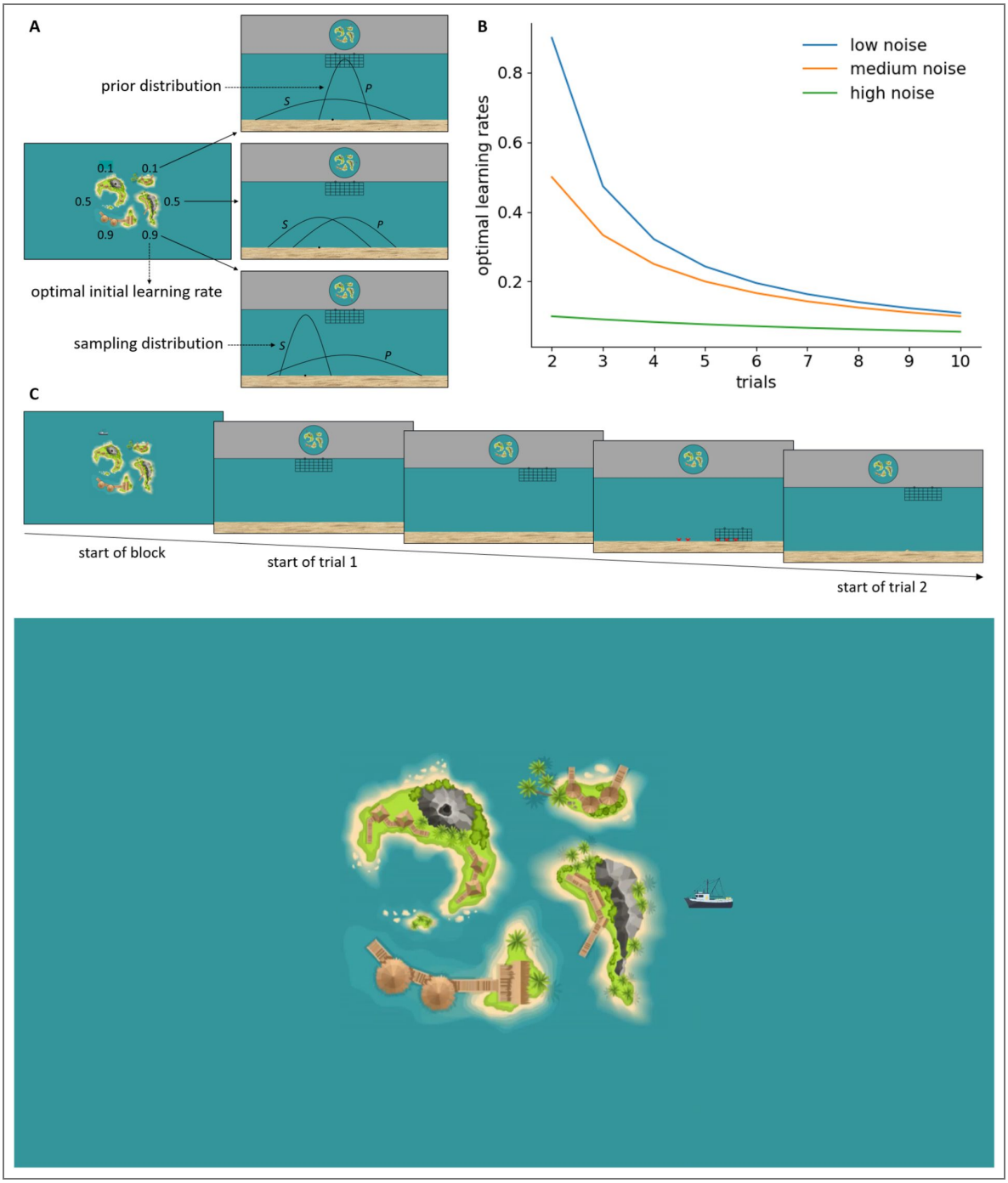


Figure 1. Experimental design.

A: Participants went fishing for crabs on six different locations around an island which differed in terms of optimal initial learning rate. At the beginning of each block, μ_s was sampled from the prior distribution $\mu_s \sim N\left(\mu_p, \sigma_p^2,\right)$, truncated between $\mu_p \pm 1.65 * \sigma_p$, with μ_p = the centre of the screen. Subsequently, on each trial, once a cage was dropped, five crabs appeared and spread out from one location in the sand sampled from the sampling distribution $S \sim N\left(\mu_s, \sigma_S^2\right)$, truncated between $\mu_s \pm 1.65 * \sigma_s$, each of which was either caught by the cage or ran away. **B:** Overview of optimal learning rates for performing the task for 1 block of trials according to the Kalman filter assuming that measurement σ_s^2 and estimate σ_p^2 on trial 1 (See *Model estimation and selection* for details). **C:** Overview of the trial procedure (see also Video 1). At the beginning of each block, participants were taken to one of six locations around the island. On each trial, participants positioned the cage somewhere along the x-axis of the screen and dropped it. As the cage sank, five crabs appeared out of one point in the sand and spread out. When the cage reached the ocean floor, crabs caught by the cage remained there while the other crabs ran away. At the start of the next trial, the cage was again at the top of the screen, but at the same x-coordinate where it was dropped in the last trial, and a little heap of sand was left where the five crabs had appeared out of the sand on the last trial.

environments were identical, which we also confirmed by analysing the feedback data. This implies that variations in the first learning rate are uniquely attributable to the meta-learned learning rates.

Behavioural results

We first evaluated whether our design was successful in inducing variations in learning rate within environments – where learning rate was defined as the percentage of the direction taken from the last cage drop to last appearance of crabs (see Methods). Specifically, consistent with the optimality analyses depicted in Figure 1B, we assumed that learning rate would decrease over the course of a block, as participants learned more and more about the crab locations within that block. Consistently, a linear mixed effects model analysis, with a random intercept for participant and random slopes for environment and trial number, indicated that learning rates significantly decreased over trials (within blocks) (Figure 2A; β = -0.056, SE = 0.006, p < .001).

Second, to probe whether people were also able to learn and adapt their learning rate to the environment-specific differences, we analysed the learning rates separately focusing on the second trial only (after their first feedback). Differences in learning rates between environments after first feedback could not be attributed to differences in experienced prediction errors, as prediction errors after the first trial were similar across the three environments. Instead, differences in this initial learning rate had to reflect a reaction to learned statistics of the environment from earlier environment visits (i.e., meta-learning). Indeed, we found that observed learning rates on the second trial of each block were significantly higher in the low noise compared to in the high noise environment, $t(49) = 2.896$, $p = .003$ (see Figure 2B). Moreover, these learning rates were also higher in the low noise compared to in the medium noise environment, $t(49) = 2.407$, $p = .01$, but not in the medium noise compared to in the high noise environment, $t(49) = 1.304$, $p = .099$.

Assuming these differences could only have developed over time, after sufficient experience with each of the environments, we also evaluated whether they were more pronounced in the second half of the experiment, compared to the first half. Figure 2C-D indeed suggests this difference in learning rates between the low and high noise environment after the second trial could not yet be observed in the first half of the experiment ($t(49) = 0.415$, $p = .34$), while it appeared in the second half ($t(49) = 2.067$, $p = .022$). However, this was not further corroborated by an interaction between time (first vs. second half of the task) and environment (low vs. medium vs. high noise environment) ($F(2, 98) = 0.923$, $p = .401$). Finally, a moving-window analysis of the 2nd-trial learning rate across blocks (with a window size of 20 blocks) did suggest that learning rates gradually decrease in the high-noise condition (see Figure 2E).

Modelling results

Next, we turned to computational modelling to evaluate which model could best describe the variations in learning rate reported above. We fitted six models to the data using hierarchical Bayesian analysis (Ahn et al., 2017). We fitted three model types, one of which assumed no local adaptations to learning rates, and two of which allowed for learning rates to be adjusted on a trial-by-trial basis. For each of these model types, we tested both a variant with (initial) learning rates for each environment separately (i.e., environment-specific), and one with a single initial learning rate (non-environment-specific). The first two models were the environment-specific and non-environment specific versions of the Rescorla-Wagner model. The Rescorla-Wagner model assumes that participants updated their μ_s estimates (within blocks) using the delta rule with a fixed learning rate (across trials, within blocks). The next two models were the environment-specific and the non-environment-specific version of the Kalman filter. The Kalman filter assumes that participants updated their μ_s estimates (within blocks) using the delta rule, where the learning rate is a function of estimation noise and measurement noise. Because estimation noise gradually decreases in a block (as people are increasingly aware of the mean crab location), learning rate gradually decreases. Here, initial estimation noise is the free, estimated parameter. The final two models were the environment-specific and non-environment-specific versions of a model that allowed for local, prediction-error weighted changes to the learning rate, here referred

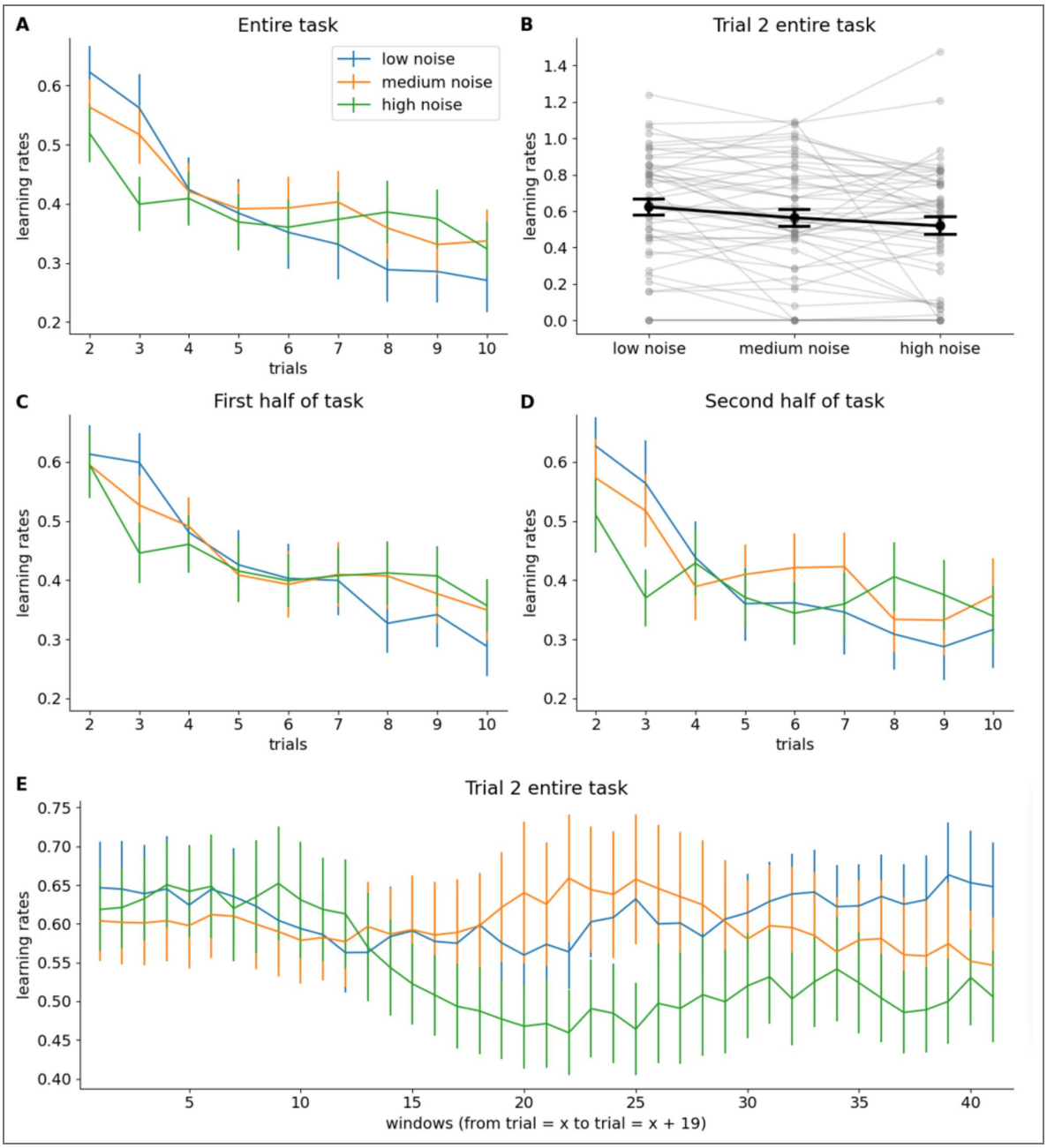


Figure 2. **Behavioural results Experiment 1.**

A: Group-level mean of each participant's median learning rate for each trial in each environment (See *Behavioural data analyses* for details). Error bars represent standard errors of the means. **B:** Detailed overview of all participants' median initial learning rates. **C-D:** Evolution of (group-level mean) learning rates over trials (within blocks) for the first half (C) and the second half (D) of the task separately. E: Moving-window analysis of trial-2 learning rate across blocks.

to as the Bai model (Bai et al., 2014; Simoens et al., 2024). The Bai model also assumes that participants updated their μ_s estimates (within blocks) using the delta rule. The Bai model further assumes that participants start with an initial learning rate that they up- and down-regulate (within blocks) in proportion to experienced prediction errors. Here, both initial learning rate and decay rate are free, estimated parameters, and both were either environment-specific or non-environment-specific.

According to the leave-one-out information criterion (LOOIC) (Vehtari et al., 2017), the environment-specific Bai model fitted the data best (Table 1), indicating that participants indeed learned to use environment-specific initial learning rates and that they indeed decreased their learning rates over trials (in contrast to what the RW model predicts), but that they did so driven by experienced prediction errors rather than in a statistically optimal way (in contrast to what the Kalman filter predicts).

The posterior probabilities that the initial learning rates (Figure 3A; estimated by the environment-specific Bai model) were higher in the low noise compared to the high noise environment was 0.999; in the low noise compared to the medium noise environment was 0.962; and in the medium noise compared to the high noise environment was 0.999. Decay rates were not significantly different from each other (Figure 3B).

Experiment 2

To establish whether our findings could be reproduced in an independent sample and to investigate where in the brain environment-specific learning rates are represented, we ran a near-exact replication of our first experiment, which 53 participants performed inside an MR-scanner. Experiment 2 consisted of 60 blocks that were identical to the blocks in Experiment 1, except that they consisted of only eight trials. Additionally, 60 blocks consisting of only two trials were randomly intermixed with the 60 longer blocks. These shorter blocks were included to increase power for analysis on the first few trials within blocks. Furthermore, we equipped the fishing boat with a laser pointer (i.e., a vertical red line from the middle of the cage to the sand) so participants could estimate more precisely where their cage would land when dropped. Similarly, to avoid miscalibrations in relation to their previous attempt, we reminded participants of their last cage location by placing a red cross wherever their cage last appeared.

Behavioural results

As in Experiment 1, we found that learning rates significantly decreased over trials within blocks (Figure 4A; β = -0.059, SE = 0.008, p < .001). Also, observed learning rates on the second trial of each block were significantly different across environments (Figure 4B; $F(2, 104) = 12.839$, p < .001). That is, they were significantly higher in the low noise than in the high noise environment ($t(52) = 3.843$, p < .001), in the low noise than in the medium noise environment ($t(52) = 1.998$, p = .025), and in the medium noise than in the high noise environment ($t(52) = 3.555$, p < .001).

Because we had twice as many first trials as in Experiment 1, we next compared quarters of the task rather than halves of the task as done in Experiment 1, as a more sensitive measure of what they learned over the course of the experiment. In line with our expectation that initial learning rates reflect learned statistics of the environment, a significant interaction between time (four quarters of the tasks) and environment (three measurement noises) indicated that learning rates on the second trial of each block became more environment-specific over time (Figure 4C-E; $F(6, 312) = 2.819$, p = .011). To unpack this interaction, in the first quarter, initial learning rates were significantly higher in the medium noise environment than in the high noise environment ($t(52) = 2.252$, p = .014), but not in the low noise compared to the high noise environment ($t(52) = 1.338$, p = .093), nor in the low noise compared to in the medium noise environment ($t(52) = -0.258$, p = .601). In the second quarter, initial learning rates were significantly higher in the low noise than in the high noise environment ($t(52) = 2.442$, p = .009) and in medium noise compared to in the high noise environment ($t(52) = 2.443$, p = .009), but not in the low noise compared to in the medium noise environment ($t(52) = - 0.12$, p = .548). In the third and fourth quarters, initial learning rates were significantly higher in the low noise compared to in the high noise environment ($t(52) = 4.164$, p <

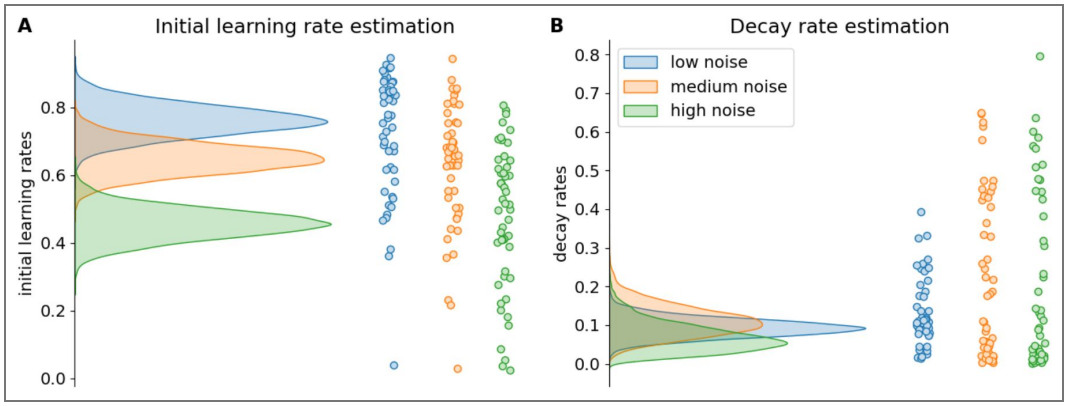
Table 1. **Model comparison**

Model	LOOIC	SE	ΔLOOIC	ΔSE
Environment-specific Bai model	28735	404	0	0
Non-environment-specific Bai model	28582	441	153	65
Environment-specific Rescorla-Wagner model	28436	398	299	67
Non-environment-specific Rescorla-Wagner model	28172	440	563	112
Environment-specific Kalman filter	27853	493	882	211
Non-environment-specific Kalman filter	27698	521	1037	211

Note. Models are ranked in descending order according to how well they fit the data. LOOIC refers to a model's approximated expected log pointwise predictive density. Higher values indicate higher out-of-sample predictive fit. SE refers to the standard error of a model's LOOIC. ΔLOOIC refers to the difference between a model's LOOIC and the top ranked model's LOOIC. ΔSE refers to the standard error of the difference between a model's LOOIC and the top ranked model's LOOIC.

Figure 3. **Bai model estimation results Experiment 1.**

The density plots on the left side of each subfigure show the full posterior densities over the means of the group-level distributions of the relevant parameters. The scatter plots on the right side of each subfigure show the means of all individual-level posterior distributions of the relevant parameters.



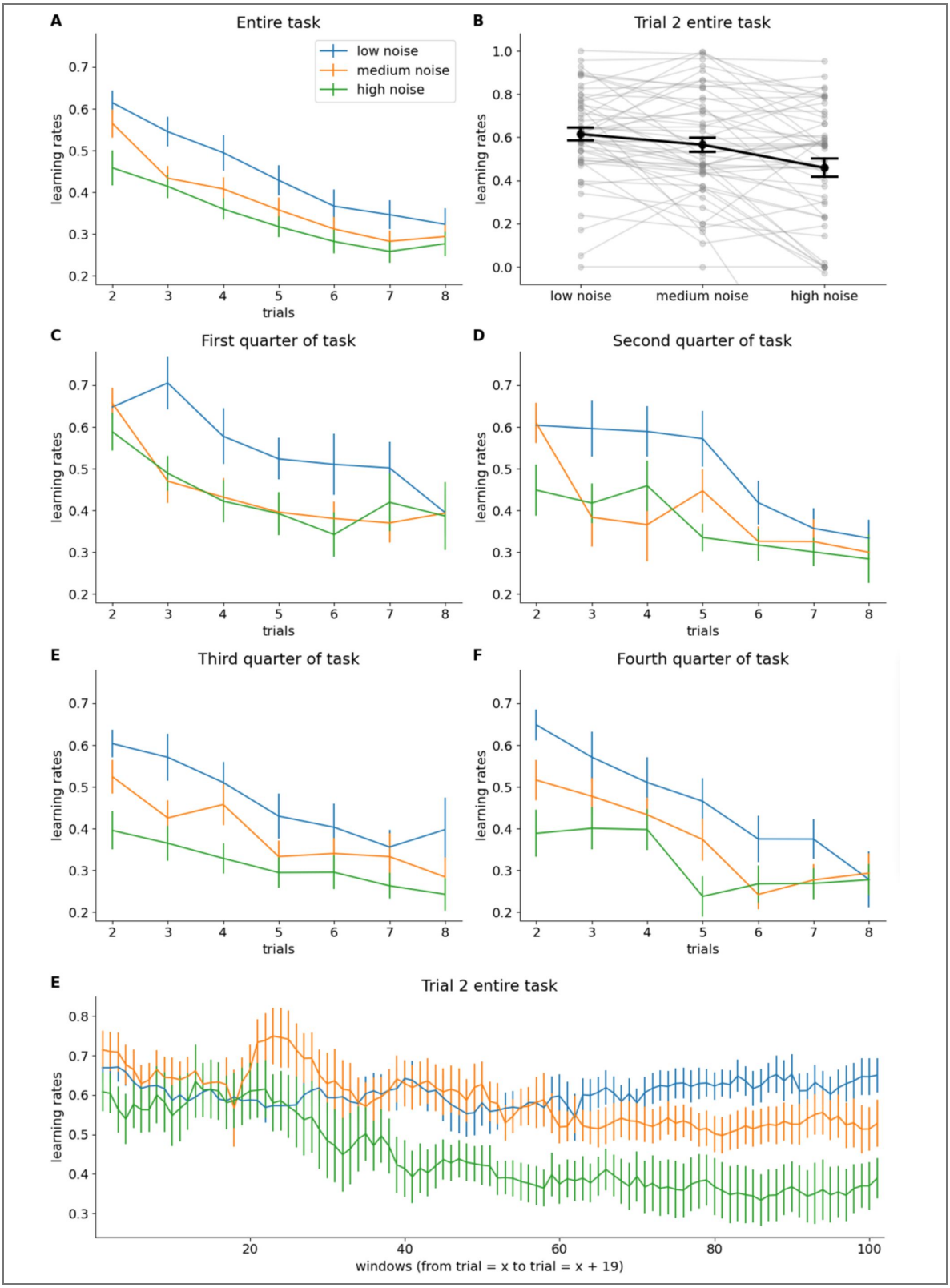


Figure 4. Behavioural results Experiment 2.

A: Group-level mean of each participant's median learning rate for each trial in each environment (See *Behavioural data analyses* for details). Error bars represent standard errors of the means. **B:** Detailed overview of all participants' median initial learning rates. One participant's median initial learning rate of -0.674 in the high noise environment is not visible on the plot. **C-F:** Evolution of (group-level mean) learning rates over trials (within blocks) for each quarter of the task separately. **G:** Moving-window analysis of 2nd-trial learning rate.

.001; $t(52) = 4.116$, $p < .001$), in the low noise compared to in the medium noise environment ($t(52) = 2.207$, $p = .016$; $t(52) = 2.829$, $p = .003$), and in the medium noise compared to in the high noise environment ($t(52) = 3.073$, $p = .002$; $t(52) = 2.425$, $p = .009$). Finally, the same moving-window analysis again suggests a gradual decrease of learning rate in the high-noise condition, and smaller decrease in the medium-noise condition (Figure 4G).

Modelling results

Replicating our findings from Experiment 1, the environment-specific Bai model fitted the data best (Table 2), indicating that participants indeed learned to use environment-specific initial learning rates and that they indeed decreased their learning rates over trials (as opposed to what the RW model predicts), but that they did so driven by experienced prediction errors rather than in a statistically optimal way (as opposed to what the Kalman filter predicts).

The posterior probabilities that initial learning rates estimated by the environment-specific Bai model (Figure 5A) were higher in the low noise compared to the high noise environment, in the low noise compared to the medium noise environment, and in the medium noise compared to the high noise environment, were 0.999, 0.992, and 0.992, respectively. Decay rates were not significantly different (Figure 5B).

fMRI results

The neural representation of environment-specific learning rates during island presentation

To investigate the neural representation of environment-specific learning rates, we first performed a whole-brain searchlight representational similarity analysis (RSA) that tested for higher similarities between locations that required the same learning rate versus locations that required different learning rates. This analysis was done on data at the time of island presentation at the start of each block, which allowed us to test whether the mere presentation of the boat informing participants of where they would be fishing for crabs next triggered a state representation that differed depending on the relevant (initial) learning rate. Our hypothesis of representations that were specific to the high, mid, and low noise environments, but not for exact location, was encoded in a corresponding learning rate model representational dissimilarity matrix (RDM; Figure 6B). We then correlated this learning rate RDM with the corresponding neural RDM throughout the whole brain for each participant and we tested which voxels showed a significant (FDR-corrected $p < .05$) correlation on the group-level. This resulted in multiple clusters of significant voxels in left as well as right OFC (Figure 6C), in accordance with our hypotheses. We also found a large cluster of significant voxels in the occipital cortex. In the next paragraph, we further interpret the activation in the two clusters.

Dissociating the representation of spatial location and learning rate during island presentation

Since differences in optimal initial learning rate and differences in spatial location between the six locations are correlated, the activity in the occipital cortex is likely driven by spatial location rather than by learned initial learning rate. Crucially, representations of initial learning rate will tend to increase over blocks because they must be learned. Indeed, the behavioural data showed that environment-specificity of initial learning rates increased over time and was most pronounced in the second half of the task. Instead, representations of spatial locations should not. This allowed us to disentangle representations of spatial location and representations of initial learning rate. We thus performed a region of interest (ROI)-based follow-up analysis within the occipital cortex (defined as the cluster of significant voxels from the whole-brain RSA). For this analysis we constructed a second model RDM that encoded the different spatial locations, rather than the environments (spatial location model RDM; Figure 6A). We then calculated how strongly each participant's neural RDM correlated with both model RDMs in the first half and the second half of the task separately. Finally, we performed a two (model RDM: spatial location vs. learning rate) by two (time: first vs. second half of the task) repeated measures ANOVA on the resulting correlations. We found a significant main effect of model RDM ($F(1, 48) = 4.732$, $p = .035$; Figure 6D), indicating that activity in the occipital cortex was indeed mostly driven by spatial

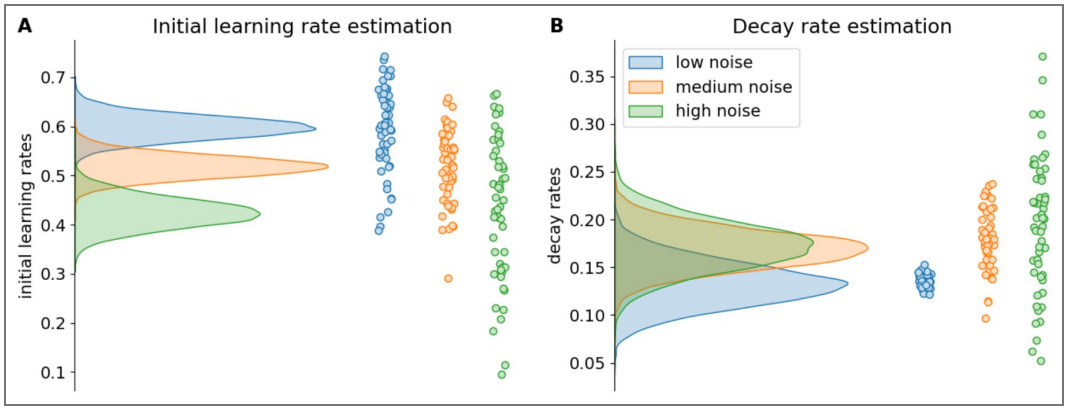
Table 2. **Model comparison**

Model	LOOIC	SE	ΔLOOIC	ΔSE
Environment-specific Bai model	34725	142	0	0
Non-environment-specific Bai model	34682	158	43	34
Environment-specific Kalman filter	34592	160	133	42
Environment-specific Rescorla-Wagner model	34567	136	158	32
Non-environment-specific Kalman filter	34523	174	202	49
Non-environment-specific Rescorla-Wagner model	34434	154	291	51

Note. Models are ranked in descending order according to how well they fit the data. LOOIC refers to a model's approximated expected log pointwise predictive density. Higher values indicate higher out-of-sample predictive fit. SE refers to the standard error of a model's LOOIC. ΔLOOIC refers to the difference between a model's LOOIC and the top ranked model's LOOIC. ΔSE refers to the standard error of the difference between a model's LOOIC and the top ranked model's LOOIC.

Figure 5. **Bai model estimation results Experiment 2.**

The density plots on the left side of each subfigure show the full posterior densities over the means of the group-level distributions of the relevant parameters. The scatter plots on the right side of each subfigure show the means of all individual-level posterior distributions of the relevant parameters.



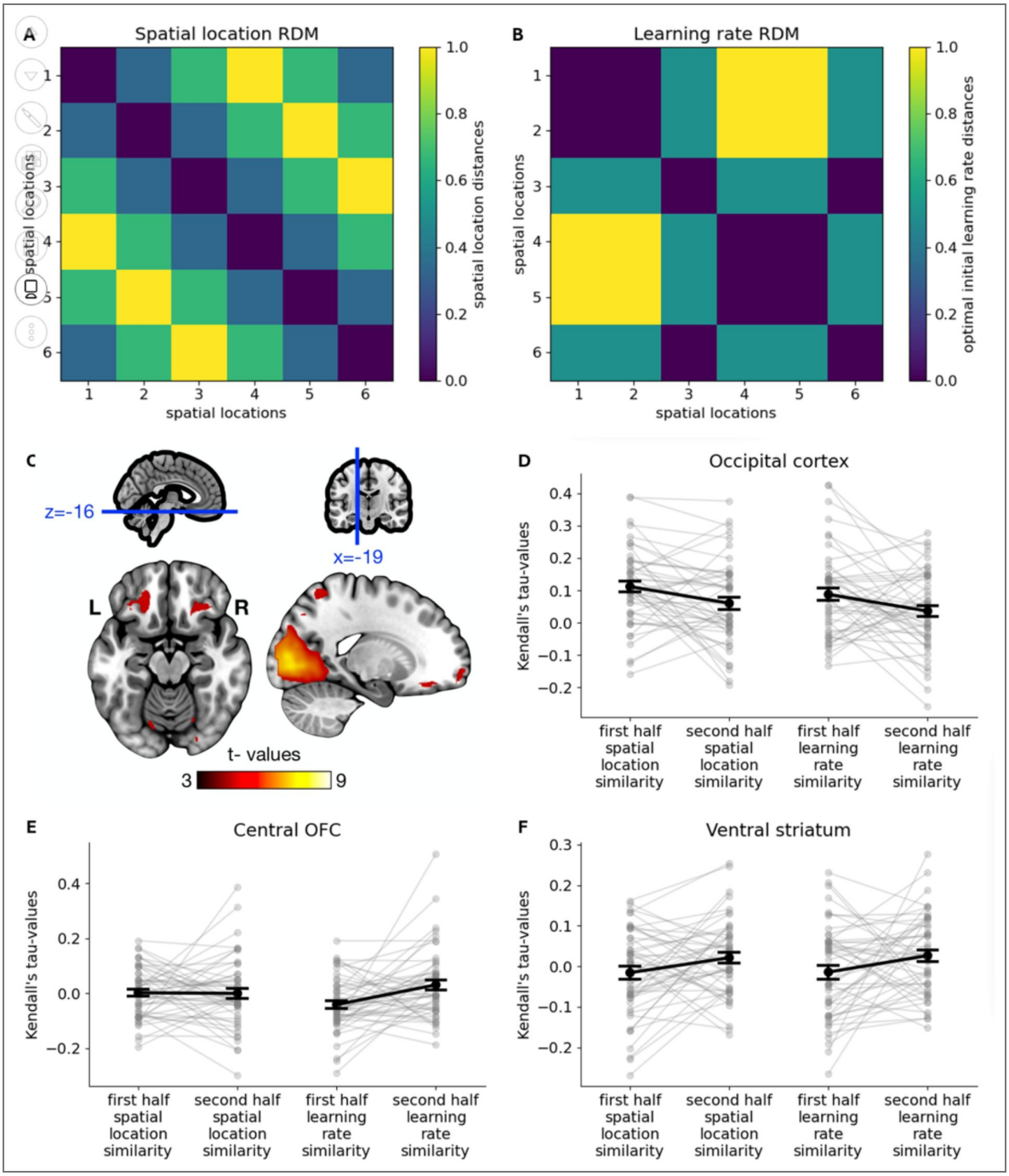


Figure 6. RSA analysis of the fMRI data.

A: spatial location RDM. **B:** Learning rate RDM. **C:** Brain map of significant t-values resulting from the whole-brain searchlight RSA of fMRI data acquired while participants had just been transported to the next location around the island (correlation with learning rate RDM). **D-F:** Interaction effect between time (first vs. second half of task) and RDM (spatial location vs. learning rate RDM) in the occipital cortex, defined as the cluster of significant voxels found in the aforementioned whole-brain searchlight RSA (D); the central OFC as defined by (Kahnt et al., 2012), based on connections to other brain regions (E); and the ventral striatum, defined as the left and right nucleus accumbens according to the AAL atlas (F). Grey dots represent individual-level Kendall's tau-values, while black dots and error bars represent group-level means and SEs of the means, respectively.

location rather than initial learning rate, as well as a significant main effect of time ($F(1, 48) = 6.665$, $p = .013$), indicating that the representation of spatial location as well as initial learning rate decreased over time. We found no interaction effect between RDM and time.

To confirm that activity in the OFC was driven by differences in initial learning rate between the six locations and to pin down where exactly in the OFC environment-specific initial learning rates were represented, we divided the OFC into six subregions as defined by (Kahnt et al., 2012), based on its connections to other brain regions. We opted for this independent ROI approach, because the significant whole-brain cluster we observed only partially overlapped with OFC, as well as with other neighbouring regions. For each of these ROIs, we then calculated how strongly each participant's neural RDM correlated with their learning rate RDM and with their spatial location RDM in the first half and the second half of the task separately. Finally, we performed a two (RDM: spatial location vs. learning rate) by two (time: first vs. second half of the task) repeated measures ANOVA on the resulting correlations. We found no main effect of RDM nor time in any of the OFC subregions, but we did find a significant interaction between RDM and time in one of the OFC subregions, namely the central OFC ($F(1, 48) = 13.076$, $p < .001$; Bonferroni corrected; Figure 6E), which is the OFC subregion that overlapped most with the largest cluster of significant voxels observed in the whole-brain RSA. Follow-up one-tailed paired t-tests confirmed that in the central OFC correlations between learning rate RDMs and neural RDMs were higher in the second half than in the first half of the task ($t(48) = 3.051$, $p = .002$), suggesting that representations of environment-specific initial learning rates increased over time. Instead, the correlation between spatial location RDMs and neural RDMs did not change across time ($t(48) = -0.109$, $p = .543$).

We also performed the same interaction analysis in the ventral striatum to study whether the ventral striatum also represented environment-specific initial learning rates during island presentation. We found no interaction effect between RDM and time, nor a main effect of RDM. However, we did find a main effect of time ($F(1, 48) = 5.499$, $p = .023$; Figure 6F), indicating that in the ventral striatum the representations of spatial location as well as initial learning rate increased over time.

Environment-sensitive neural processing of prediction errors during crab fishing

Finally, we also investigated how participants processed reward prediction errors. That is, we used the distance between the cage location (their prediction) and target location (the reward) as an approximation of reward prediction errors. To investigate neural activity during this phase we focused on the neural response to these prediction errors after the first and second trial, which was included as a parametric modulator in a series of first-level general linear models (GLMs). As a first analysis, we performed a whole brain (univariate) analysis on this parametric modulator, averaged over all environments, to evaluate whether we observed a typical response to prediction errors in the brain (Figure 7A-B). Indeed, in line with previous studies on reward prediction error processing (Calderon et al., 2021; O'Doherty et al., 2004; Pessiglione et al., 2006; Schultz et al., 1997), we observed significant clusters of voxels in the left and the right striatum, including in the ventral striatum.

Next, we investigated whether there were interactions between environment (low vs. medium vs. high noise environment) and time (first vs. second half of the task) on these neural prediction error signals. Importantly, as mentioned above, our measure is only an (inverse) approximation of reward prediction error, because it measures the mismatch in location but not how much uncertainty people had around their point estimation. For example, seeing crabs appear on the exact cage location could still come with varying degrees of positive surprise depending on the uncertainty around their prediction, making it also difficult to estimate where exactly the prediction error turned form negative to positive. However, a region that is sensitive to the reward information in this signal, should show a neural response that is further dependent on environment and time, reflecting people's ability to learn over time the more reward-informative nature of this signal in low noise environments. While we anticipated that the ventral striatum's response would be modulated by the learning context, it is important to note that we held no specific a priori hypotheses regarding the direction of these effects.

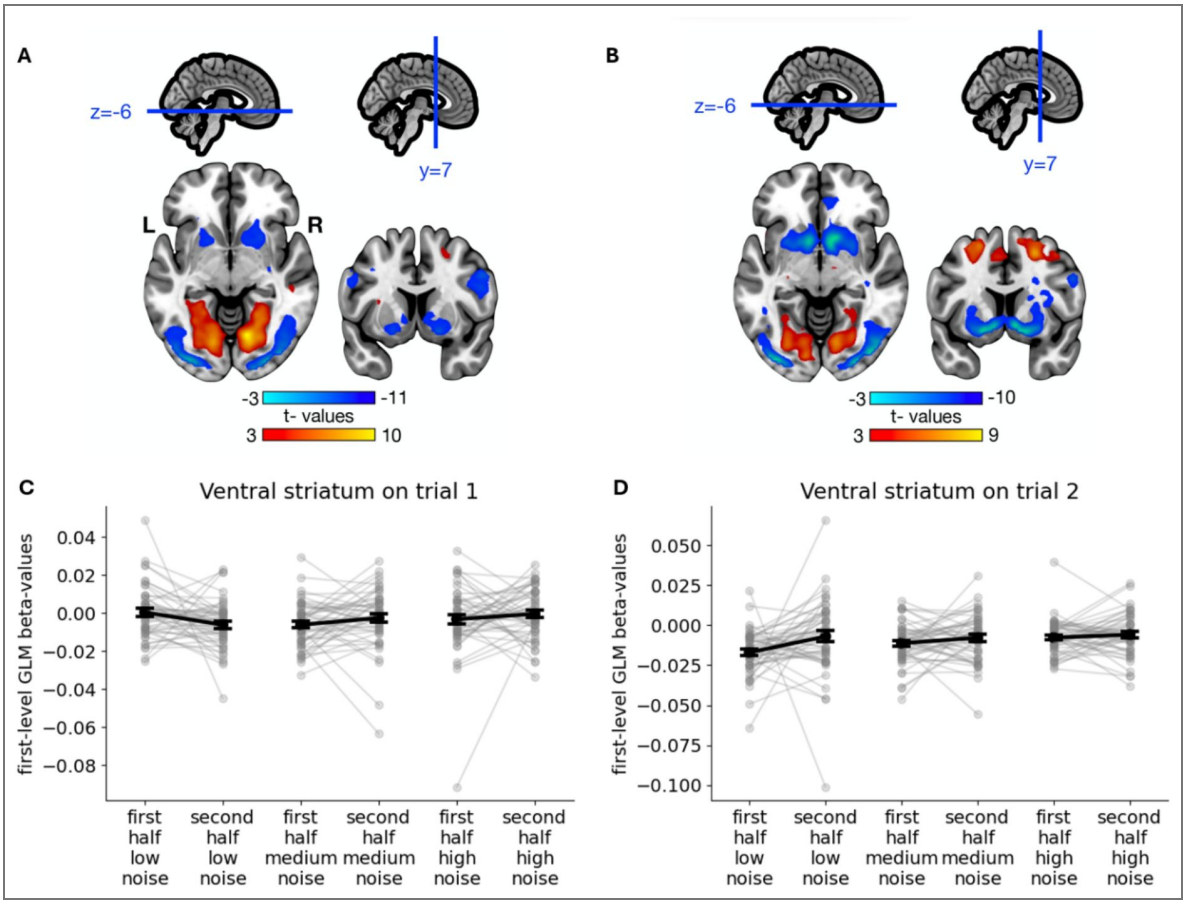


Figure 7. Results of the analyses of the effect of prediction error on the fMRI data.

A-B: Brain map of significant t-values resulting from the whole-brain (univariate) tests of voxel activity being (parametrically) modulated by prediction error on trial 1 (A) and trial 2 (B). **C-D:** Interaction effect between time (first vs. second half of task) and environment (low vs. medium vs. high measurement noise) on the modulating effect of prediction error on trial 1 (C) and trial 2 (D) on ventral striatum activity. This ROI was defined as the left and right nucleus accumbens according to the AAL atlas. Grey dots represent individual-level GLM beta-values, while black dots and error bars represent group-level means and SEs of the means, respectively.

We investigated this in the six OFC subregions and the ventral striatum using repeated measures ANOVAs. There were no significant effects in any of the OFC subregions. The ventral striatum, however, showed a significant interaction effect between environment and time ($F(1, 48) = 4.222$, $p = .017$; Figure 7C) in an analysis focusing on feedback processing during the first trial. Follow-up t-tests confirmed that the ventral striatum responded differently to prediction errors in the low noise environment in the second half compared to the first half of the task ($t(48) = 2.284$, $p = .027$), while it did not show such an evolution in the medium noise environment ($t(48) = -1.468$, $p = .149$), nor the high noise environment ($t(48) = -0.842$, $p = .404$). Specifically, ventral striatum activity showed a more negative response to larger location prediction errors. This pattern is consistent with its documented role in encoding reward prediction errors (Calderon et al., 2021). Namely, being closer to the centre of where the crabs appeared (i.e. smaller location prediction errors) corresponds to outcomes that are less negative or more positive than expected (i.e. smaller negative or larger positive reward prediction errors). On the second trial, there were no significant effects in any of the OFC subregions, while the ventral striatum only showed a significant main effect of time ($F(1, 48) = 4.884$, $p = .032$; Figure 7D), suggesting it may have become less sensitive to reward prediction errors on the second trial over time.

Discussion

Humans can adapt how they learn in response to environmental demands, but how these adaptations unfold over time, and how to dissociate different types of adjustments, remains poorly understood. Here, we developed a new paradigm that systematically disentangles two types of learning rate adaptations: a fast and transient response to local prediction errors, and a slower form of meta-learning an optimal learning rate as a function of higher-level environmental statistics. Specifically, we designed a gamified task in which participants fish for crabs on six different fishing locations around an island. The environmental statistics about the crabs' hiding spots implied different optimal learning rates for the different locations. We extracted participants' learning rates on each trial, which allowed us to test both (1) whether participants dynamically adjusted their learning rates in response to locally experienced prediction errors and (2) whether we could observe, above and beyond these local adaptations, different initial learning rates tailored to the environmental statistics. Across two experiments, we observed that participants did both: They immediately adapted their learning rates on a trial-by-trial basis in response to just-experienced prediction errors, but also learned, over time, to use different initial learning rates on the different locations around the island.

Computational modelling confirmed our findings. We fitted six models to the data, and in both experiments the best fitting model was the environment-specific Bai model, which implemented both the learning of environment-specific initial learning rates (across blocks) and learning rate updating proportional to recently experienced prediction errors (within blocks). While the Kalman filter provides optimal learning rates for the present task on every trial based on underlying environment statistics, the Bai model assumes a (learned) initial learning rate which is up- or downregulated by experienced prediction errors. According to the Kalman filter, learning rates should quickly decrease towards 0 irrespective of the initial learning rate, and independently of (individual differences in sensitivity to) experienced prediction errors. While participants' learning rates do decrease over trials, they stabilise around 0.3, indicating that people are more responsive to noise than is optimal.

Overall, our behavioural data and modelling suggest that, over the course of both experiments, participants learned to instantaneously retrieve relevant learning rates upon arrival on any given location around the island. Our findings go beyond previous studies that documented changes in learning rates over time (Behrens et al., 2007; Browning et al., 2015; Goris et al., 2021), by showing that people can also switch back and forth between learning rates across environments (see also (Simoens et al., 2024)). As such, the present study provides support for recent theories of meta-learning (Botvinick et al., 2009; Holroyd & Verguts, 2021; Silvetti et al., 2018; Wang et al., 2018), as well as cognitive control (Abrahamse et al., 2016; Braem et al., 2019; Chiu & Egner, 2017), which posit that cognitive control is implemented as the environment-specific

regulation of task execution parameters, such as learning rate. Interestingly, our work is also consistent with that of others who have shown that also the local adaptation strategy in itself, may reflect an environment-specific strategy based on prior experience with these environments' generative structure (Bakst & McGuire, 2021, 2021; Lee et al., 2020).

We also investigated which brain regions represent sustained, meta-learned associations between environments and learning rates, enabling one to instantaneously retrieve the relevant learning rate when revisiting an environment. Importantly, previous studies examined neural correlates of learning rates during outcome evaluation, where learning rates may be adjusted online as a function of locally experienced prediction errors (e.g., (Behrens et al., 2007; Browning et al., 2015; Nassar et al., 2012). In contrast, our RSA analysis targeted neural activity at island presentation, before any outcome information was available. At this moment, learning rates cannot be updated based on current feedback and instead reflected the retrieval of a previously learned, environment-specific learning-rate settings. This difference reflects our hypothesis that the OFC represents the latent states in a cognitive map of the task (Knudsen & Wallis, 2022; Moneta et al., 2024; Schuck et al., 2018; Wilson et al., 2014), which are expected to activate as soon as the agents can infer which task state it is in. Several studies have identified such "partially observable" task states in the medial OFC (Bradfield et al., 2015; Schuck et al., 2016; Tan et al., 2025; Wimmer & Büchel, 2019), in line with the region identified here (but see e.g., (Ongur & Price, 2000), for important anatomical distinctions between medial and lateral OFC and (Tan et al., 2025) for an example of related functions in lateral OFC). Our finding extends this notion by suggesting a link between OFC and meta learning, wherein meta-learned information becomes encapsulated in task states (Hattori et al., 2023; Moneta et al., 2024). Consistently, OFC has been shown to represent task states (Moneta et al., 2024; Stalnaker et al., 2015; Wilson et al., 2014). While earlier evidence shows that the OFC represents concrete aspects of task states, such as task-relevant stimulus features (Schuck et al., 2016), we hypothesized that the OFC also represents more abstract aspects, such as learned, environment-specific learning rates. Indeed, we showed that the central OFC gradually came to represent these environment-specific learning rates (or the environment-specific statistics that drive them). While previous work speculated that these different levels could have different neural underpinnings (Sharpe et al., 2019), our findings indicate OFC might signal states on multiple levels. This does not imply identical learning dynamics; fast-changing trial-specific states might be learned through activity dynamics, while higher-level contextual states could involve synaptic plasticity.

We further observed that the ventral striatum learned to differentially respond more to positive reward prediction errors (or less to negative reward prediction errors) depending on the currently relevant environment-specific learning rate. Similar to previous studies, the ventral striatum responded more to prediction errors where the noise was lowest (Diederen et al., 2016) but see (Mah et al., 2024). This heightened sensitivity of the ventral striatum to reward prediction errors in low-noise environments aligns with the fact that these signals are most behaviorally informative in stable contexts, where a higher learning rate is optimal for guiding future behavior. The ventral striatum also became less sensitive to the second trial's prediction error over time. Presumably, as participants gained more experience with the task's global reward structure (specifically that all targets center around a fixed mean), the first reward prediction error per round became the primary source of information, rendering subsequent reward prediction errors within that round exponentially less behaviorally relevant over time. We also saw a prediction error signal in the ACC, but only on the second trial. Interestingly, prediction errors on the second trial are the first events after participants have formed an initial, local estimate of the crab's location and prediction errors could meaningfully signal the need to update internal models or control settings (Hayden et al., 2011; Silvetti et al., 2018), in line with earlier studies suggesting the ACC's role in uncertainty-driven belief updating in this task (McGuire et al., 2014). Taken together, our fMRI data analyses suggest that the OFC, more specifically the central OFC, represents environment-specific learning rates, which may in turn affect how the ventral striatum and ACC responded to prediction errors.

Our findings are in line with recent conceptualizations of learning across two time scales, both in artificial and biological (Binz et al., 2024) agents. Here, fast learning would presumably occur in (neural) activation space, whereas slower learning would occur in weight space. Although our study did not allow specifying the locus of (fast and slow) learning, a direct comparison between artificial and biological agents was reported in (Hattori et al., 2023). They used a two-armed bandit task to study how mice as well as deep reinforcement learning models adapt their behaviour over time. They found that, in mice, the OFC is crucial for slow, across-session learning, which gradually refines neural circuits that support fast, within-session learning. The researchers also observed parallels between the mechanisms underlying learning at these two different time scales in mice and deep reinforcement learning models. That is, both systems employed synaptic plasticity mechanisms to shape neural connectivity for learning on the slow time scale, but instead on recurrent activity dynamics for learning on the fast time scale (Duan et al., 2017; Wang et al., 2018). Indeed, blocking synaptic plasticity in the OFC disrupted across-session learning, but left within-session learning intact in expert mice (Hattori et al., 2023). Thus, slow, across-session meta-learning may involve plasticity-based mechanisms in the OFC, which serve to improve fast, within-session learning of new tasks through recurrent activity dynamics. Here, we provided first evidence that a similar dual process may operate in humans.

A recent theme in artificial intelligence and computational neuroscience is that (artificial and biological) agents (should) learn to cluster the environments they are confronted with; and associate different (low- or high-level) parameters to each such environment (Collins & Frank, 2013; Verbelen et al., 2022). This approach leads to efficient learning, not least because it shields against catastrophic interference. Here, we used only three environments (albeit on six locations), so the clustering was relatively easy in our case. However, future studies could use a design similar to the one presented here to investigate the neural underpinnings of clustering environments (and associated learning rates) in a continuous range of environments around the island. Similarly, the present study focused on differentiating between environments in terms of learning rate. However, human as well as nonhuman agents are often confronted with novel situations, in which they may instantaneously deduce appropriate settings for task execution parameters such as learning rate, across contexts; in brief, adaptive agents can generalize task execution parameters across similar contexts. Future studies could leverage the task developed for the present study to investigate the neural underpinnings of this generalisation of (abstract) knowledge by, for example, introducing more locations later on in the task.

In conclusion, the present study demonstrates the importance of differentiating between two time scales of adaptation in human learning rates. Fast time scale learning within environments and slow time scale learning about environments likely involve different cognitive processes with different neural underpinnings. Nevertheless, this distinction has thus far been largely overlooked, in favour of the fast time scale. Future research could leverage the experimental design presented here to further investigate the slow time scale as well. Our gamified design (Allen et al., 2024) can also provide new ways to test theories about the relation between meta-learning and development (Nussenbaum & Hartley, 2024) or psychological pathologies, such as theories of autism which posit that autism is related to deficits in the detection of environmental differences in learning opportunities (Goris et al., 2021; van de Cruys et al., 2014).

Methods

Participants

50 participants (42 female, 8 male) were recruited for Experiment 1, and 53 participants (41 female, 12 male) for Experiment 2, through Sona (*https://www.ugent.sona-systems.com/*). All participants were between 18 and 35 years old. Because all participants caught about the same number of crabs, no participants were excluded from the behavioural data analyses. Since no response time deadline was implemented, no trials were excluded from the analyses. However, two participants were excluded from the fMRI data analyses because they were left handed, and another two because of technical problems with the fMRI data acquisition.

Experiment 1 was approved by the Ghent University Psychology and Educational Sciences Ethical Committee, and Experiment 2 by the Ghent University Medical Ethical Committee. Participants signed informed consents prior to participation. Experiment 1 took participants about 45 minutes to complete, in return for which they received a participation fee of €10. Participants in Experiment 2 received a participation fee of €35, because we also administered fMRI recording. In both experiments, the participant who caught the most crabs, received a €50 gift certificate for *bol.com* (an online store that offers general merchandising products). All data and code can be found on *https://osf.io/qft2p* for Experiment 1, and on *https://osf.io/be4td* for Experiment 2.

Experimental design

In Experiment 1, participants performed a novel crab-fishing task, which was programmed in jsPsych (de Leeuw et al., 2023). At the start of each of 60 blocks, a boat took them to one of six locations around an island (Figure 1C). There they dropped a cage ten times (trials) trying to catch crabs. Each time a cage was dropped, one location was sampled from the sampling distribution $S \sim N\left(\mu_s, \sigma_s^2\right)$, truncated between $\mu_s \pm 1.65 * \sigma_s$ in order to avoid confusing participants with the occasional extreme outlier as well as off-screen locations. Before the cage reached the ocean floor, five crabs appeared and spread out evenly from this location, each of which was either caught by the cage or ran away. At the beginning of each block, μ_s was sampled from the prior distribution $\mu_s \sim N\left(\mu_p, \sigma_p^2\right)$, truncated between $\mu_p \pm 1.65 * \sigma_p$, with μ_p = the centre of the screen.

Crucially, the standard deviations of both prior and sampling distributions were dependent on the location around the island. On two randomly selected adjacent locations, σ_p was large (18.75% of the screen width), while σ_s was small (6.25% of screen width), making this a low noise environment. Here, a high initial learning rate is optimal since the mean of the sampling distribution could be far away from the centre of the screen, but all crabs will cluster close together. Thus, in estimating the mean of the sampling distribution, it makes sense to give a lot of weight to the first crabs (i.e., use a high learning rate), and exponentially decrease the learning rate afterwards. On the two adjacent locations on the exact opposite side of the island, the situation was reversed, making this a high noise environment. Here, a low (initial) learning rate is optimal since the mean of the sampling distribution can only be near the centre of the screen, but crabs can appear far away from each other (and the centre of the screen). Thus, in estimating the mean of the sampling distribution, it makes sense not to give too much weight to any individual crabs (i.e., use a low learning rate). Finally, on the two locations in between, the standard deviations σ_p and σ_s were intermediate (12.5% of screen width) and equal to each other, making this a medium noise environment and requiring an intermediate (initial) learning rate.

Participants visited all six locations around the island once in randomised order before visiting all locations a second time in (re)randomised order, and so on. The sailing of the boat from the previous to the next location unfolded over three to five seconds, depending on how far apart the locations were (one second stationary at previous location, followed by 60 degrees per second of sailing around the island, followed by one second stationary at the next location). Next, participants could move the cage to the left using the f-key and to the right using the j-key. Tapping the key would move the cage 1% of the screen width in the corresponding direction, while holding the key would slide the cage in that direction more quickly. There was no response time deadline. Participants could drop the cage by pressing the space bar, after which feedback unfolded over the course of 1.5 seconds. For the first 500 ms the cage sank until halfway to the bottom of the ocean; for the next 500 ms the cage sank all the way to the bottom of the ocean while five crabs appeared out of one point on the ocean floor and spread out to cover the same proportion of the screen width as the cage (18.75%); for the last 500 ms crabs that were not caught by the cage ran away (see Figure 1C), while crabs that were caught by the cage remained in place. The usage of a wide cage as well as five crabs was to ensure that participants could still catch some crabs in the high noise environment. At the start of each trial after the first trial (within blocks), the cage was moved back to the top of the screen at the x-coordinate where it was dropped on the last trial. The heap of sand where five crabs had crawled out of the sand on the last trial, would still be visible in

order to help participants determine where to drop the cage next. While participants were fishing for crabs, a radar at the top of the screen reminded them of where around the island they were at all times.

Participants were instructed that around this island, crabs live in groups that are denser near the centre than towards the edges and that the local group of crabs would not change location while they were fishing on its location. The local group of crabs would only change location while they were fishing somewhere else. Hence, they should try to drop their cage over the centre of the group to maximise reward. They were also instructed that these groups of crabs might have different sizes around the island, so that they should keep track of where around the island they are.

After receiving the instructions, participants first performed a short practice phase during which they performed one block in each of the three environments. During the practice phase, participants could not see where around the island they were, while the (normally distributed) group of crabs under the sand was made visible while they were fishing.

During the actual task, the group of crabs was, of course, not visible while participants were fishing for crabs. However, they did receive block feedback at the end of each block. That is, at the end of each block, the group of crabs was made visible, as well as all ten (heaps of sand at) locations where crabs had crawled out of the sand during the block. During the fishing, only the location where crabs had appeared after the previous cage was dropped, was made visible as a little heap of sand. Finally, the task was divided into four rounds, in between which participants could take a short break.

Experiment 2 was programmed in PsychoPy (Peirce, 2007) and consisted of 60 blocks that were identical to the blocks in Experiment 1, except that they consisted of only eight trials. Additionally, 60 blocks that consisted of only two trials were randomly intermixed with the 60 longer blocks. These blocks were included to increase power for analyses on the very first (i.e., the most informative) trials within blocks. At the end of these shorter blocks, participants received no block feedback.

To make the design suitable for the MR-scanner, some additional minor changes were made. At the start of a block, the boat did not sail to a new location around the island but was immediately presented at the new location for a duration of three to seven seconds. Next, participants could move the cage to the left and to the right using their right index finger and right middle finger, respectively, on a response box that was placed in their right hand. Participants could drop the cage using their left index finger on a response box placed in their left hand. A laser pointer was added to the cage that pointed straight down from the centre of the cage to the ocean floor. Feedback unfolded over the course of 750 ms. During the first 250 ms, the cage sank until halfway to the bottom of the ocean. During the next 250 ms, the cage sank further to the bottom of the ocean and five crabs appeared out of one point on the ocean floor and spread out to cover the same proportion of the screen width as the cage. Before the start of the last 250 ms, crabs that were not caught by the cage disappeared from the screen. After the first and second trial of each block, there was an intertrial interval of three to seven seconds during which only the heap of sand where crabs had just crawled out of the ocean floor as well as a red cross (also on the ocean floor) at the centre of where the cage had just been dropped remained visible on the screen. Finally, block feedback lasted for three to seven seconds.

Behavioural data analysis

All data and analysis scripts can be found on the Open Science Framework. Participants' continuous responses allowed us to solve the delta rule:

$$E_t = E_{t-1} + \alpha_t \left(M_{t-1} - E_{t-1}\right)$$

where E_t and $E_{t\text{-}1}$ are the participant's estimates of μ_s on trials t and t - 1, respectively (i.e., the locations where the participant dropped the cages), α_t is the participant's learning rate on

trial t, and M_t is the location where the five crabs appeared on trial t, for the learning rate α for each trial $t > 1$. That is, each trial $t > 1$ we calculated:

$$\alpha_t = \frac{E_t - E_{t-1}}{M_{t-1} - E_{t-1}}$$

This allowed us to test whether participants showed a decrease in learning rate over the course of a block, as they learned more and more about the location of crabs within that block. To this end, we performed a linear mixed effects model analysis, with a random intercept for participant and a random slope for environment as well as for trial number (as well as for the environment-trial interaction), on participants' median learning rates (for each environment-trial combination). We opted for median rather than mean learning rates (on a participant-level) because we observed some extreme outlier learning rates on trials directly following a trial in which crabs appeared directly next to where participants dropped their last cage. Here, any movement of the cage has an effect on the resulting learning rate which is unlikely to be proportional to participants' actual learning rates (by blowing up the numerator, either positively or negatively, in the above formula for α_t.

To investigate whether participants' median learning rates on the second trial of each block were significantly different in the three environments and whether participants gradually learned to use different initial learning rates in the different environments, we divided Experiment 1 into two halves and Experiment 2 into four quarters (i.e., the four functional runs in the MR-scanner) and performed a two (time: the two halves) or four (time: the four runs) by three (environment: the three environments) repeated measures ANOVA on participants' median learning rates on the second trial of each block. We interpreted the significant main effect of environment and the significant interaction effect between time and environment using one-tailed paired t-tests. We opted for median rather than mean learning rates (on a participant-level) because we observed some extreme outlier learning rates on trials directly following a trial in which crabs appeared directly next to where participants dropped their last cage. Here, any movement of the cage has an effect on the resulting learning rate which is unlikely to be proportional to participants' actual learning rates (because the denominator in the formula for α_t is close to zero and the equation hence unstable).

Model estimation and selection

We fitted six models to the data using hierarchical Bayesian analyses (HBA) (Ahn et al., 2017). The HBA was performed in Stan (Carpenter et al., 2017), which uses Hamiltonian Monte Carlo (HMC) sampling, a variant of Markov chain Monte Carlo (MCMC) sampling. The first two models were the environment-specific and non-environment specific versions of the Kalman filter. The Kalman filter assumes that people update their estimates using the delta rule, with a learning rate that depends on both estimate and measurement uncertainty:

$$\alpha_t = \frac{\sigma_{p,t}^2}{\sigma_{p,t}^2 + \sigma_s^2}$$

where α_t is the participant's current learning rate; $\sigma_{p,t}^2$ is the participant's current estimate uncertainty, which initially is the variance of the prior distribution; and σ_S^2 is the measurement uncertainty, which here is the variance of the sampling distribution. Crucially, learning rate decreases over time because:

$$\sigma_{p,t}^2 = (1 - \alpha_{t-1})\,\sigma_{p,t-1}^2$$

Initial estimate uncertainty $\sigma_{p,t=0}^2$ and measurement uncertainty σ_S^2 cannot both be free parameters, because they trade off against each other, which would result in bimodal and unreliable posteriors (Daw et al., 2006). Therefore, we fixed measurement uncertainty to 0.125^2 (the true value in the medium noise environment) and only estimated initial estimate uncertainty. As an alternative procedure, we also tried fixing measurement uncertainty to the true value in each environment separately. The two approaches resulted in different posterior densities for estimate uncertainties, but in similar posterior densities for learning rates. Moreover, both approaches resulted in almost identical values for the LOOIC. We

conclude that we can safely fix measurement uncertainty to 0.125^2 across environments. In the environment-specific version of the model, each participant was assumed to use separate initial estimate uncertainties for each environment, while in the non-environment-specific version of the model, participants were assumed to use the same initial estimate uncertainty in both environments.

The next two models were the environment-specific and non-environment specific versions of the Rescorla-Wagner model, which assumes that people update their estimates according to the delta rule (with a constant learning rate). In the environment-specific version of the model, each participant was assumed to use separate learning rates for each environment, while in the non-environment-specific version of the model, participants were assumed to use the same learning rate in both environments.

The final two models were the environment-specific and non-environment specific versions of the Bai model (Bai et al., 2014), which is in the general family of models that adapt learning rate as a function of prediction errors (Krugel et al., 2009; Pearce & Hall, 1980). Hence, the Bai model also supposes that people update their estimates according to the delta rule, but use a learning rate dependent on recently experienced prediction errors:

$$\alpha_t = \eta |M_{t-1} - E_{t-1}| + (1 - \eta)\alpha_{t-1}$$

where η is a decay rate, which also takes a value between 0 and 1 and determines how much learning rates are affected by recently experienced prediction errors. That is, when prediction errors are large, learning rate increases, but when prediction errors are small, learning rate decreases exponentially. In the environment-specific version of the model, each participant was assumed to use separate initial learning rates and decay rates for each environment, while in the non-environment-specific version of the model, participants were assumed to use the same learning rate and decay rate in both environments.

For each environment (low vs. medium vs. high noise), individual-level free parameters were assumed to be drawn from a group-level normal distribution specific to that environment. For the means of these group-level distributions, we used uniform priors between 0 and 1, while for the standard deviations we used half-Cauchy (0, 5) priors. For the individual-level parameters we also used bounded uniform priors. To minimize the dependence between the means and standard deviations of group-level distributions, we used non-centred parameterisations. To maximise the efficiency of HMC sampling, parameters were first estimated in an unbounded space and then probit-transformed to the relevant bounded space (Ahn et al., 2017).

For each model, 4000 samples were drawn from the posterior distributions, the first 1000 of which were discarded as burn-in, across four sampling chains, resulting in a total of 12,000 posterior samples. Convergence of posterior distributions was checked by visually inspecting the traces and by numerically checking the Gelman-Rubin statistics (Gelman & Rubin, 1992), which were all well below 1.1, for each estimated parameter.

We used the LOOIC (Vehtari et al., 2017) for model comparisons. The LOOIC approximates the log pointwise posterior predictive density of observed data, which is the out-of-sample predictive accuracy of a model. Therefore, higher values indicate higher out-of-sample predictive fit of the model to the data.

To calculate the posterior probability that learning rates were higher in one environment than in another, we calculated the proportion of posterior samples in which the group-level mean learning rate was higher in the one environment than in the other. Thus, a posterior probability higher than 95% corresponds to a one-tailed p-value lower than 0.05.

Parameter recovery

To make sure that the experimental design and model estimation procedure would allow for the reliable estimation of the model parameters, we performed parameter recovery simulations. Specifically, using the aforementioned design details, we simulated 16 datasets with each of the three models in each of the three environments separately. For the RW model we used each of the

learning rates {0.2, 0.4, 0.6, 0.8} four times as the true mean of the group-level distribution of this parameter. For the Kalman filter we used each of the estimate uncertainties {0.25, 0.5, 0.75, 1} four times as the true mean of the group-level distribution of this parameter. For the Bai model we used each combination of learning rates {0.2, 0.4, 0.6, 0.8} and decay rates {0.2, 0.4, 0.6, 0.8} as the true mean of the group-level distribution of these parameters. For each parameter, we then randomly sampled 50 true parameter values from the group-level normal distribution with the relevant mean and an SD of 0.2 for learning rates and decay rates, and 0.25 for estimate uncertainties. We then simulated a dataset using these parameter values and the experimental design described above and fitted the model to this simulated dataset using our model estimation procedure. Finally, for each model and each environment separately, we calculated parameter recovery rates by correlating all (16 datasets x 50 participants = 800) true parameters to the corresponding estimated parameters, that is, the individual-level posterior means. These simulations indicated that the model could be reliably fitted to our behavioural data. For learning rates in the RW model, estimate uncertainties in the Kalman filter, and initial learning rates in the Bai model recovery rates were higher than 0.975 in each environment. For decay rates in the Bai model parameter recovery rates were 0.807, 0.868 and 0.918 for the low, medium and high noise environment, respectively.

Model recovery

To ensure that the experimental design and model selection procedure would allow for the reliable selection of the model fitting the data best, we also performed model recovery simulations. Specifically, using the experimental design described above, we simulated 10 datasets with each of the six models described above. For each model parameter, we used the estimated posterior mean and SD of the group-level distribution (of the empirical data) as the mean and SD, respectively, of the group-level distribution (of the simulated data), and randomly sampled 50 values from this distribution. For each model, we then simulated a dataset using these parameter values and the experiment design described above and fitted all six models to this simulated dataset using the model estimation procedure described above. Finally, using the model selection procedure described above, we tested for each simulated dataset which model fitted it best and calculated model recovery rates as the proportion of datasets simulated with each of the three models that was best fit by the correct model. Model recovery rates were 100% for each of the six models.

Model validation

For validating the winning model, we used the posterior predictive check method (Gelman et al., 1996). This method takes participants' fitted model parameters and uses them to simulate responses given their individual trial sequence. Simulated and true response patterns can then be compared to determine how well the model captures participants' behaviour (Palminteri et al., 2017). Specifically, for each participant, we used the means of the participant-level posterior densities over the relevant parameters to simulate a response sequence conditional on the trial sequence this participant had received. We then (Pearson) correlated each participant's true response sequence with the simulated response sequence.

Posterior predictive checks indicated that the environment-specific Bai model indeed adequately captured participants' behaviour. The average correlation between participants' true choice sequences and the response sequences obtained by simulating data using their parameter estimates was 0.793 in Experiment 1 and 0.863 in Experiment 2. More specifically, the average correlations were 0.927, 0.81 and 0.642 for the low, medium and high noise environment, respectively, in Experiment 1 and 0.955, 0.879 and 0.754 for the low, medium and high noise environment, respectively, in Experiment 2. Although still medium to large and reliable, we do note that these correlations were slightly lower in the high noise environment. In this environment, individual outcomes are considerably less indicative of the latent mean, which may reduce the usefulness of the trial-by-trial, prediction-error–driven learning-rate adjustments that we see in the low and medium noise environments. Under extreme conditions of variability,

people may rely less on delta-rule updating and more on alternative strategies (d'Acremont & Bossaerts, 2016; Reynders et al., 2026), such as exploratory adjustments or heuristics that are not explicitly captured by the Bai model, but also outside the scope of the present paper.

fMRI data acquisition and preprocessing

T1-weighted MPRAGE structural images (1 mm isotropic voxels, 256 x 256 matrix, 176 axial slices, 9° flip angle), GRE field map images (528 ms TR, 7.38 TE, 60° flip angle), and T2*-weighted EPI functional data (2.5 mm isotropic voxels, 64 x 64 matrix, 1780 ms TR, 27 ms TE, 66° flip angle) were acquired on a 3 T Prisma scanner system (Siemens) with a 64 channel head coil. Functional data were acquired in 4 runs, each of which lasted about 12 minutes.

MRI data were preprocessed using fMRIPrep 23.1.0 (Esteban et al., 2019) and involved motion correction, field map based geometric undistortion, slice timing correction, coregistration of anatomical and functional scans, normalization into MNI space, and spatial smoothing with a 5 mm FWHM Gaussian kernel.

fMRI data analyses

First-level (subject-wise) general linear modelling was done in SPM (*https://www.fil.ion.ucl.ac.uk/spm/*) and involved regressors of interest that captured stimulus onset events (using an event-related design (i.e., with all event durations = 0); see below) and nuisance regressors that reflected participant movement (six regressors) as well as the global signal (one regressor). All regressors were convolved with a canonical hemodynamic response function.

Standard GLMs were used to estimate voxel activations associated with stimulus display. First-level models included separate regressors for each of the seven events of interest in each experimental block crossed with each of the six locations around the island plus the above described seven nuisance regressors for each run separately. The seven events of interest (regressor onset times) were: (1) presentation of the island at the beginning of each block, (2) the onset of the first trial, (3) the end of the feedback of the first trial, (4) the onset of the second trial, (5) the end of the feedback of the second trial, (6) the onsets of each remaining trial in the experimental block (when it was a long block), and (7) the onset of the block feedback (when it was a long block). We also included (mean-centred absolute) prediction error (in the current trial) as a parametric modulator for events 3 and 5 separately (as well as for each location around the island separately). This resulted in 244 whole brain maps of parameter estimates ("betas"; ((seven events of interest + two parametric modulators) * six locations around the island + seven nuisance regressors) * four runs).

The RSA focused on the beta maps capturing voxel activation during the presentation of the island at the beginning of each block. Before proceeding to the RSA, we normalized each voxel to its own within-run mean, by subtracting the voxel's overall mean activation from each location's activation, so that the overall mean is zero (Diedrichsen & Kriegeskorte, 2017). Next, we performed a whole-brain searchlight RSA in three steps. Firstly, for each participant, we constructed a 24×24 design-based RDM with as rows and columns the six locations in each run (out of 4) and in each cell the dissimilarity between the optimal initial learning rate in the row's location and the optimal initial learning rate in the column's location (which could be 0, 1, or 2). Secondly, for each participant, we went through the entire brain in spherical searchlights with a radius of three voxels to construct 24×24 neural RDMs and compare (the upper triangle of) these RDMs to (the upper triangle of) the design-based RDM (also excluding within run cells) by computing Kendall's Tau. Within each searchlight, the neural dissimilarity between each pair of locations was computed as one minus the Pearson correlation between the voxel wise activations for those locations. This resulted in a brain map of dissimilarity values (tau-values) for each participant where each voxel's tau-value is computed with that voxel in the centre of a searchlight. Finally, we used one-tailed t-tests to test which voxels had tau-values significantly higher than

zero, which indicates that they responded differently to different locations around the island, using a significance threshold of 0.05 corrected for multiple comparisons using the false discovery rate (FDR).

Next, we performed a ROI-based follow-up analysis with the occipital cortex, defined as the cluster of significant voxels from the whole brain searchlight RSA, as ROI in order to tease apart representations of spatial location and initial learning rate in this ROI, which were strongly correlated with each other. For this analysis we constructed a second 24×24 design-based RDM with as rows and columns the six locations in each of the four runs, and in each cell the dissimilarity between the spatial location in the row's location and the spatial location in the column's location (which could be 0, 1, 2, or 3). We then correlated the parts of (the upper triangle of) both design-based RDMs that concerned the first two runs (excluding within run cells) to the corresponding part of the neural RDMs, and then did the same for the last two runs. Finally, we performed a two (design feature: spatial location vs. optimal initial learning rate) by two (time: first vs. second half of the task) repeated measures ANOVA on the resulting correlations.

Next, we performed more in depth RSAs within predefined anatomical ROIs within the OFC to test whether within these regions the representation of initial learning rate level gradually became stronger than the representation of spatial location. To this end, we performed four RSAs within each ROI. That is, we correlated the parts of (the upper triangle of) both design-based RDMs that concerned the first two runs (excluding within run cells) to the corresponding part of the neural RDMs, and then did the same for the last two runs. Finally, we performed a two (design feature: spatial location vs. optimal initial learning rate) by two (time: first vs. second half of the task) repeated measures ANOVA on the resulting correlations. Where a significant interaction effect was found, we used one-tailed paired t-tests to interpret this interaction effect. The ROIs were created using SPM's wfupickatlas toolbox. The OFC subregions were defined as in, based on connections to other brain regions. Here, we defined multiple ROIs within the OFC based on what is known about its structure, rather than defining an ROI based on the cluster of significant voxels observed in the whole brain RSA described above, as we did for the occipital cortex, because we wanted to check whether there were no significant effects in ROIs in which clusters of voxels did not exceed the stringent significance threshold applied in the whole brain RSA described above. Since we tested six ROIs, we used Bonferroni correction for multiple comparisons, which lowered the significance threshold to .05/6 = .008.

Although we found no significant clusters of voxels in the ventral striatum in the whole-brain RSA described above, we also performed the same repeated measures ANOVA we performed in the occipital cortex and the OFC in the ventral striatum to check if there was no effect there that did not survive the stringent significance threshold applied in the whole-brain RSA (or was cancelled out by an interaction between time and RDM, considering the whole-brain RSA exclusively tests for a main effect of RDM). This ROI was created using SPM's wfupickatlas toolbox and was defined as the left and right nucleus accumbens according to the AAL atlas.

We also performed all of the (whole-brain and ROI-based) RSAs described above on neural data acquired during the presentation of feedback after the first two trials in each block.

Finally, we also checked the effect of prediction errors on brain activity using univariate analyses. First, as a sanity check, we used a whole-brain approach in which we averaged together (over blocks and environments) all whole-brain maps of first level beta-values for the parametric modulator "prediction error" on the regressor "end of feedback presentation after the first trial" and checked which voxels had significant beta-values using FDR correction for multiple comparisons. Crucially, we also tested whether there was a significant interaction effect between time (first vs. second half of the task) and environment (low vs. medium vs. high noise environment) on the parametric modulator "prediction error" on the regressor "end of feedback presentation after the first trial" in one of the seven ROIs described above (six OFC subregions and the ventral striatum) using repeated measures ANOVAs and, if applicable, follow-up paired one-tailed t-test to interpret significant interaction effects. We used the same approach to investigate the effect of prediction errors on brain activity during feedback presentation on the second trial of each block.

For analysing the reward localiser, we used the same approach as for analysing the main task up to and including first-level analyses, where we used the first three events of interest from the main task. This resulted in 13 whole brain maps of parameter estimates (betas; three events of interest * two locations around the island + seven nuisance regressors). Next, we simply took the univariate contrast between the high and the low reward conditions at the first event of interest (i.e., island presentation). We did this because this should have shown us which brain regions were responsive to expected reward during the event of interest that our main analyses were focused on. In that way, we could control for this effect there. However, no brain regions were significantly responsive to reward expectancy during our reward localiser.

Data availability

All behavioural and neuroimaging data as well as the code for the analyses of the data presented in this paper are publicly available at *https://osf.io/be4td/*.

Acknowledgements

The authors thank Matthew Nassar for useful discussion.

Additional information

Funding

This work was funded by an FWO fellowship awarded to Jonas Simoens (#11K5121N), an FWO project grant awarded to Tom Verguts and Senne Braem (G010319N), and an ERC Starting grant awarded to Senne Braem (European Union's Horizon 2020 research and innovation program, Grant agreement 852570). NWS was supported by the Excellence Strategy of the Federal Government and the Länder and a ERC Starting Grant (ERC StG REPLAY-852669). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

Jonas Simoens: data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualisation, writing - original draft preparation. Senne Braem and Tom Verguts: conceptualisation, funding acquisition, methodology, project administration, supervision, writing - review & editing. Nicolas Schuck: supervision, writing - review & editing. Pieter Verbeke, Haopeng Chen, Stefania Mattioni, and Mengqiao Chai: data curation, investigation.

Funding

Funder	Grant reference number	Author
Fonds Wetenschappelijk Onderzoek (FWO)	11K5121N	Stefania Mattioni Haopeng Chen Nicolas W Schuck Tom Verguts Mengqiao Chai Pieter Verbeke
Fonds Wetenschappelijk Onderzoek (FWO)	G010319N	Pieter Verbeke
EC | European Research Council (ERC)	https://doi.org/10.3030/852570	Senne Braem
EC | European Research Council (ERC)	https://doi.org/10.3030/852669	Senne Braem

Author ORCID iDs

Senne Braem: https://orcid.org/0000-0002-2619-8225
Stefania Mattioni: https://orcid.org/0000-0001-8279-6118
Nicolas W Schuck: https://orcid.org/0000-0002-0150-8776
Tom Verguts: https://orcid.org/0000-0002-7783-4754

Additional files

Video 1

References

Abrahamse E. L., Braem S., Notebaert W., Verguts T. (2016) Grounding cognitive control in associative learning. *Psychological Bulletin* **142**:693-728 https://doi.org/10.1037/bul0000047 | PubMed

Ahn W.-Y., Haines N., Zhang L. (2017) Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. *Computational Psychiatry* **1**:24-57 https://doi.org/10.1162/CPSY_a_00002 | PubMed

Allen K., Brändle F., Botvinick M., Fan J. E., Gershman S. J., Gopnik A., Griffiths T. L., Hartshorne J. K., Hauser T. U., Ho M. K., *et al.* (2024) Using games to understand the mind. *Nature Human Behaviour* **8**:1035-1043 https://doi.org/10.1038/s41562-024-01878-9 | PubMed

Bai Y., Katahira K., Ohira H. (2014) Dual learning processes underlying human decision-making in reversal learning tasks: Functional significance and evidence from the model fit to human behavior. *Frontiers in Psychology* **5**:1-8 https://doi.org/10.3389/fpsyg.2014.00871 | PubMed

Bakst L., McGuire J. T. (2021) Eye movements reflect adaptive predictions and predictive precision. *Journal of Experimental Psychology: General* **150**:915-929 https://doi.org/10.1037/xge0000977 | PubMed

Behrens T. E. J., Woolrich M. W., Walton M. E., Rushworth M. F. S. (2007) Learning the value of information in an uncertain world. *Nature Neuroscience* **10**:1214-1221 https://doi.org/10.1038/nn1954 | PubMed

Binz M., Dasgupta I., Jagadish A., Botvinick M., Wang J. X., Schulz E. (2024) Meta-learned models of cognition. *Behavioral and Brain Sciences* https://doi.org/10.1017/s0140525x23003266 | PubMed

Botvinick M. M., Niv Y., Barto A. G. (2009) Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition* **113**:262-280 https://doi.org/10.1016/j.cognition.2008.08.011 | PubMed

Bradfield L. A., Dezfouli A., van Holstein M., Chieng B., Balleine B. W. (2015) Medial Orbitofrontal Cortex Mediates Outcome Retrieval in Partially Observable Task Situations. *Neuron* **88**:1268-1280 https://doi.org/10.1016/j.neuron.2015.10.044 | PubMed

Braem S., Bugg J. M., Schmidt J. R., Crump M. J. C., Weissman D. H., Notebaert W., Egner T. (2019) Measuring Adaptive Control in Conflict Tasks. *Trends in Cognitive Sciences* **23**:769-783 https://doi.org/10.1016/j.tics.2019.07.002 | PubMed

Browning M., Behrens T. E., Jocham G., O'Reilly J. X., Bishop S. J. (2015) Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nat Neurosci* **18**:590-596 https://doi.org/10.1038/nn.3961 | PubMed

Calderon C. B., De Loof E., Ergo K., Snoeck A., Boehler C. N., Verguts T. (2021) Signed Reward Prediction Errors in the Ventral Striatum Drive Episodic Memory. *Journal of Neuroscience* **41**:1716-1726 https://doi.org/10.1523/jneurosci.1785-20.2020 | PubMed

Carpenter B., Gelman A., Hoffman M. D., Lee D., Goodrich B., Betancourt M., Brubaker M., Guo J., Li P., Riddell A. (2017) Stan: A probabilistic programming language. *Journal of Statistical Software* **76**:1-32 https://doi.org/10.18637/jss.v076.i01 | PubMed

Chiu Y.-C., Egner T. (2017) Cueing cognitive flexibility: Item-specific learning of switch readiness. *Journal of Experimental Psychology: Human Perception and Performance* **43**:1950-1960 https://doi.org/10.1037/xhp0000420 | PubMed

Collins A. G. E., Frank M. J. (2013) Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review* **120**:190-229 https://doi.org/10.1037/a0030852 | PubMed

Cook J. L., Swart J. C., Frobo M. I., Geurts D. E. M., Den Ouden H. E. M. (2019) Catecholaminergic modulation of meta-learning. *eLife* **8**:e51439 https://doi.org/10.7554/elife.51439 | PubMed

d'Acremont M., Bossaerts P. (2016) Neural Mechanisms Behind Identification of Leptokurtic Noise and Adaptive Behavioral Response. *Cerebral Cortex* **26**:1818-1830 https://doi.org/10.1093/cercor/bhw013 | PubMed

Daw N. D., Doherty J. P. O., Dayan P., Seymour B., Dolan R. J. (2006) Cortical substrates for exploratory decisions in humans. *Nature* **441**:876-879 https://doi.org/10.1038/nature04766 | PubMed

Dayan P., Kakade S., Montague P. R. (2000) Learning and selective attention. *Nature Neuroscience* **3**:1218-1223 https://doi.org/10.1038/81504 | PubMed

de Leeuw J. R., Gilbert R. A., Luchterhandt B. (2023) jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments. *Journal of Open Source Software* **8**:5351 https://doi.org/10.21105/joss.05351

Diederen K. M. J., Spencer T., Vestergaard M. D., Fletcher P. C., Diederen K. M. J., Spencer T., Vestergaard M. D., Fletcher P. C., Schultz W. (2016) Adaptive Prediction Error Coding in the Human Midbrain and Striatum Facilitates Behavioral Adaptation and Learning Efficiency Adaptive Prediction. *Neuron* **90**:1127-1138 https://doi.org/10.1016/j.neuron.2016.04.019 | PubMed

Diedrichsen J., Kriegeskorte N. (2017) Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology* **13**:e1005508 https://doi.org/10.1371/journal.pcbi.1005508 | PubMed

Duan Y., Schulman J., Chen X., Bartlett P. L., Sutskever I., Abbeel P. (2017) RL2: Fast reinforcement learning via slow reinforcement learning. *arXiv* https://doi.org/10.48550/arXiv.1611.02779

Esteban O., Markiewicz C. J., Blair R. W., Moodie C. A., Isik A. I., Erramuzpe A., Kent J. D., Goncalves M., Dupre E., Snyder M., *et al.* (2019) fmriPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods* **16** https://doi.org/10.1038/s41592-018-0235-4 | PubMed

Gelman A., Rubin D. B. (1992) Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* **7**:457-472 https://doi.org/10.1214/ss/1177011136

Goris J., Silvetti M., Verguts T., Wiersema J. R., Brass M., Braem S. (2021) Autistic traits are related to worse performance in a volatile reward learning task despite adaptive learning rates. *Autism* **25**:440-451 https://doi.org/10.1177/1362361320962237 | PubMed

Hattori R., Hedrick N. G., Jain A., Chen S., You H., Hattori M., Choi J.-H., Lim B. K., Yasuda R., Komiyama T. (2023) Meta-reinforcement learning via orbitofrontal cortex. *Nature Neuroscience* **26**:2182-2191 https://doi.org/10.1038/s41593-023-01485-3 | PubMed

Hayden B. Y., Heilbronner S. R., Pearson J. M., Platt M. L. (2011) Surprise signals in anterior cingulate cortex: Neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience* **31**:4178-4187 https://doi.org/10.1523/JNEUROSCI.4652-10.2011 | PubMed

Holroyd C. B., Verguts T. (2021) The best laid plans: Computational principles of ACC. *Trends in Cognitive Sciences* **25** https://doi.org/10.1016/j.tics.2021.01.008 | PubMed

Kahnt T., Chang L. J., Park S. Q., Heinzle J., Haynes J.-D. (2012) Connectivity-Based Parcellation of the Human Orbitofrontal Cortex. *Journal of Neuroscience* **32**:6240-6250 https://doi.org/10.1523/JNEUROSCI.0257-12.2012 | PubMed

Kingma D. P., Ba J. (2017) Adam: A Method for Stochastic Optimization. *arXiv* https://doi.org/10.48550/arXiv.1412.6980

Knudsen E. B., Wallis J. D. (2022) Taking stock of value in the orbitofrontal cortex. *Nature Reviews Neuroscience* **23**:428-438 https://doi.org/10.1038/s41583-022-00589-2 | PubMed

Krugel L. K., Biele G., Mohr P. N. C., Li S.-C., Heekeren H. R. (2009) Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proceedings of the National Academy of Sciences of the United States of America* **106**:17951-17956 https://doi.org/10.1073/pnas.0905191106 | PubMed

Lee S., Gold J. I., Kable J. W. (2020) The human as delta-rule learner. *Decision* **7**:55-66 https://doi.org/10.1037/dec0000112

Mah A., Golden C. E. M., Constantinople C. M. (2024) Dopamine transients encode reward prediction errors independent of learning rates. *Cell Reports* **43**:114840 https://doi.org/10.1016/j.celrep.2024.114840 | PubMed

Mathys C. (2011) A Bayesian Foundation for Individual Learning Under Uncertainty. *Front Hum Neurosci* **5**:39-39 https://doi.org/10.3389/fnhum.2011.00039 | PubMed

McGuire J. T., Nassar M. R., Gold J. I., Kable J. W. (2014) Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. *Neuron* **84**:870-881 https://doi.org/10.1016/j.neuron.2014.10.013 | PubMed

Moneta N., Grossman S., Schuck N. W. (2024) Representational spaces in orbitofrontal and ventromedial prefrontal cortex: Task states, values, and beyond. *Trends in Neurosciences* **47**:1055-1069 https://doi.org/10.1016/j.tins.2024.10.005 | PubMed

Nassar M. R., Rumsey K. M., Wilson R. C., Parikh K., Heasly B., Gold J. I. (2012) Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience* **15**:1040-1045 https://doi.org/10.1038/nn.3130 | PubMed

Nussenbaum K., Hartley C. A. (2024) Understanding the development of reward learning through the lens of meta-learning. *Nature Reviews Psychology* **3**:424-438 https://doi.org/10.1038/s44159-024-00304-1

O'Doherty J. P., Dayan P., Schultz J., Deichmann R., Friston K. J., Dolan R. J. (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**:452-454 https://doi.org/10.1126/science.1094285 | PubMed

Ongur D., Price C. J. (2000) The Organization of Networks within the Orbital and Medial Prefrontal Cortex of Rats, Monkeys and Humans. *Cerebral Cortex* **10**:206-219 https://doi.org/10.1093/cercor/10.3.206 | PubMed

Palminteri S., Wyart V., Koechlin E. (2017) The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences* **21**:425-433 https://doi.org/10.1016/j.tics.2017.03.011 | PubMed

Pearce J. M., Hall G. (1980) A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review* **87**:532-552 https://doi.org/10.1037/0033-295x.87.6.532 | PubMed

Peirce J. W. (2007) PsychoPy—Psychophysics software in Python. *Journal of neuroscience methods* **162**:8-13 https://doi.org/10.1016/j.jneumeth.2006.11.017 | PubMed

Pessiglione M., Seymour B., Flandin G., Dolan R. J., Frith C. D. (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* **442**:1042-1045 https://doi.org/10.1038/nature05051 | PubMed

Rescorla R. A., Wagner A. (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black A., Prokasy W. (Eds). *Classical Conditioning II: Current Research and Theory* Appleton Century Crofts. pp. 64-99

Reynders J., Verguts T., Braem S. (2026) Strategic variability in humans, pigeons, and rats. *Psychological Review* https://doi.org/10.1037/rev0000620 | PubMed

Schuck N. W., Cai M. B., Wilson R. C., Niv Y. (2016) Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron* **91**:1402-1412 https://doi.org/10.1016/j.neuron.2016.08.019 | PubMed

Schuck N. W., Wilson R., Niv Y. (2018) A State Representation for Reinforcement Learning and Decision-Making in the Orbitofrontal Cortex. In: Morris R, Bornstein A, Shenhav A (Eds). *Goal-Directed Decision Making* Elsevier. pp. 259-278 https://doi.org/10.1016/B978-0-12-812098-9.00012-7

Schultz W., Dayan P., Montague P. R. (1997) A neural substrate of prediction and reward. *Science* **275**:1593-1599 https://doi.org/10.1126/science.275.5306.1593 | PubMed

Schweighofer N., Doya K. (2003) Meta-learning in Reinforcement Learning. *Neural Networks* **16**:5-9 https://doi.org/10.1016/S0893-6080(02)00228-9 | PubMed

Sharpe M. J., Stalnaker T., Schuck N. W., Killcross S., Schoenbaum G., Niv Y. (2019) An Integrated Model of Action Selection: Distinct Modes of Cortical Control of Striatal Decision Making. *Annual Review of Psychology* **70**:53-76 https://doi.org/10.1146/annurev-psych-010418-102824 | PubMed

Silvetti M., Vassena E., Abrahamse E. L., Verguts T. (2018) Dorsal anterior cingulate-brainstem ensemble as a reinforcement meta-learner. *PLOS Computational Biology* https://doi.org/10.1371/journal.pcbi.1006370 | PubMed

Simoens J., Verguts T., Braem S. (2024) Learning environment-specific learning rates. *PLOS Computational Biology* https://doi.org/10.1371/journal.pcbi.1011978 | PubMed

Stalnaker T. A., Cooch N. K., Schoenbaum G. (2015) What the orbitofrontal cortex does not do. *Nature Neuroscience* **18**:620-627 https://doi.org/10.1038/nn.3982 | PubMed

Sutton R. S., Barto A. G. (2018) *Reinforcement Learning: An Introduction* MIT Press.

Tan L., Qiu Y., Qiu L., Lin S., Li J., Liao J., Zhang Y., Zou W., Huang R. (2025) The medial and lateral orbitofrontal cortex jointly represent the cognitive map of task space. *Communications Biology* **8**:163 https://doi.org/10.1038/s42003-025-07588-w | PubMed

van de Cruys S., Evers K., Hallen R. V. D., Eylen L. V., Boets B., Wagemans J. (2014) Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review* **121**:649-675 https://doi.org/10.1037/a0037665 | PubMed

Vehtari A., Gelman A., Gabry J. (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27**:1413-1432 https://doi.org/10.1007/s11222-016-9696-4

Verbeke P., Verguts T. (2024) Humans adaptively select different computational strategies in different learning environments. *Psychological Review* https://doi.org/10.1037/rev0000474 | PubMed

Verbelen T., Tinguy D. D., Mazzaglia P., Çatal O., Safron A. (2022) Chunking Space and Time with Information Geometry. In: NeurIPS. pp. 1-6

Wang J. X., Kurth-nelson Z., Kumaran D., Tirumala D., Soyer H., Leibo J. Z., Hassabis D., Botvinick M. M. (2018) Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience* **21**:860-868 https://doi.org/10.1038/s41593-018-0147-8 | PubMed

Wilson R. C., Takahashi Y. K., Schoenbaum G., Niv Y. (2014) Orbitofrontal Cortex as a Cognitive Map of Task Space. *Neuron* 267-279 https://doi.org/10.1016/j.neuron.2013.11.005 | PubMed

Wimmer G. E., Büchel C. (2019) Learning of distant state predictions by the orbitofrontal cortex in humans. *Nature Communications* **10**:2554 https://doi.org/10.1038/s41467-019-10597-z | PubMed

Peer reviews

Reviewer #1 (Public review):

Summary:

Simoens and colleagues use a continuous estimation task to disentangle learning rate adjustments on shorter and longer timescales. They show that participants rapidly decrease learning rates within a block of trials in a given "location", but that they also adjust learning rates for the very first trial based on information accrued gradually about the statistics of each location, which can be viewed as a form of metalearning. The authors show that the metalearned learning rates are represented in patterns of neural activity in the orbitofrontal cortex, and that prediction errors are represented in a constellation of brain regions including ventral striatum, where they are modulated by expectations about error magnitude to some degree. The work opens the door to future work focusing on how exactly these signals contribute to adaptive behavior.

Strengths:

The authors build on an interesting task design allowing them to distinguish moment-to-moment adjustments in learning rate from slower adjustments in learning rate corresponding to slowly gained knowledge about the statistics of specific "locations". Behavior and computational modeling clearly demonstrate that individuals adjust to environmental statistics in a sort of metalearning. fMRI data reveal representations of interest including those related to adjusted learning rates and their impact on the degree of prediction error encoding in the striatum.

Weaknesses:

It was nice to see that the authors could distinguish differences between the OFC signals that they observed and those in the visual regions based on changes through the session. However, the linkage between these brain activations and a functional role in generating behavior remains somewhat unclear, opening the door for alternative interpretations.

Comments on revised version.

I appreciate the authors responses and they have largely addressed my concerns. I understand the concerns about power with regard to the individual differences/behavioral analyses included in the rebuttal. However, my personal view, which is perhaps a matter of taste, is that the paper would benefit from a description of these results - along with a clear description of why the authors are hesitant to draw a strong interpretation from the negative result.

https://doi.org/10.7554/eLife.108223.2.sa2

Reviewer #2 (Public review):

Summary:

Across two experiments, this work presents a novel spatial predictive inference paradigm that facilitates the investigation of meta-learning across multiple environments with distinct statistics, as well as more local learning from sequences of observations within an environment. The authors present behavioral data indicating that people can indeed learn to distinguish between noise levels and calibrate their learning rates accordingly across environments, even on initial trials when revisiting an environment. They complement their behavioral results with computational modeling, further bolstering claims of both local and global adaptation. Additional fMRI results support the role of OFC in this meta-learning process, with central OFC activity reflecting similarity between environments. This similarity emerges over time with task experience. Holistically, this paradigm and these data add to our understanding of how humans dynamically adapt their behavior on different timescales.

Strengths:

The novel paradigm represents a clever and creative expansion of spatial predictive inference tasks. The cover story was well chosen to facilitate an intuitive understanding of both the differences between environments, and the estimation of the mean within environments.

Additionally, the authors present complementary results from two experiments, which strengthens the behavioral findings. This is especially effective as the initial experiment's results were a bit noisy, and the modifications within the second experiment increased both power and the specificity/accuracy of participant predictions. Taken together, the behavioral results provide convincing evidence that participants did distinguish environments based on their underlying statistics and adapted their initial behavior accordingly.

Beyond this, the combination of behavioral results, computational modeling, and neuroimaging enhances the impact of the work. It paints a fuller picture of whether and how humans meta-learn the global statistics of environments, and this is an important direction for the field of adaptive learning.

Weaknesses:

Throughout much of the paper, the authors refer to the distinctions between environments primarily as differences in "initial learning rates" or "environment-specific learning rates." The optimal initial learning rate did indeed differ across environments -- the result of differences in underlying task statistics. These differences in task statistics result in distinct optimal initial learning rates and also vary with aspects of spatial position (e.g. vertical position in the example figure). The authors convincingly show that OFC activity increasingly reflects these variables throughout task experience. Given that these variables vary together, future work will be needed to distinguish whether particular variables drive these dynamics, or whether together they combine to evoke the representational differences.

The current work is also quite suggestive of meaningful individual differences in both local and global adaptive learning, in line with other prior work on predictive inference. This is perhaps underexplored in this data set, but certainly leaves the topic ripe for follow up going forward.

Finally, more information on all clusters that survived multiple comparisons correction would be useful, even in the absence of a priori hypotheses. For instance, there is commentary in the discussion section on the ACC, but this is not mentioned in the results, and it is unclear whether there were other undescribed clusters that survived correction.

https://doi.org/10.7554/eLife.108223.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public review):***
>
> *It was nice to see that the authors could distinguish differences between the OFC signals that they observed and those in the visual regions based on changes through the session. However, the linkage between these brain activations and a functional role in generating behavior was left unexplored. Without further exploration, it is hard to tell exactly what role the signals might be playing, if any, in the behavior of interest.*

To link the behavioral with the fMRI data, we now correlated fMRI decoding accuracy with behavioral performance. We studied behavioral performance in two ways: the difference in high versus low noise environment learning rates, and mean accuracy (i.e., absolute prediction error). We correlated both measures with the decodability of the environment in the central OFC. Each correlation was calculated either in the full experiment, or only the second half. However, none of these correlations were significant (all $p > .1$). Given the difficulty of interpreting this result, and our lack of statistical power for doing individual difference analyses, we decided not to report these analyses in the final paper.

> ***Reviewer 2 (public review):***
>
> *(1) The authors make the distinction between meta-learned "global" learning rates and within environment learning rate adaptation in response to "local" fluctuations/observations. Though the experimental paradigm is novel, there are certainly links to prior work - for instance, though change point structures don't entail revisiting unique environments, they do require meta-learning from environmental statistics that is distinct from transient local adaptation to prediction errors. This tendency to increase one's learning rate after large prediction errors is appropriate in change point environments, though, as is true in this study, the amount of increase should be dependent on. This represents a similar kind of slower-timescale learning or reuse of more "global" parameters, and can be seen to different extents in prior work. It might benefit readers if the authors were to link the current work to previous research more explicitly to draw clearer connections between the approaches and findings.*

We thank the reviewer for their very helpful literature suggestions and now contextualize and discuss our findings in light of relevant literature.

> *(2) Throughout much of the paper, the authors refer to the distinctions between environments primarily as differences in "initial learning rates" or "environment-specific learning rates." This is particularly prominent when discussing fMRI results. Though the optimal initial learning rate did differ across environments, this was the result of differences in underlying task statistics. It will be important to clarify this throughout the text, because of the confounds between task statistics and initial learning rate (and to some extent, the position on the screen), it is not possible to separate the impact of these specific variables. This is also relevant to understanding the justification for using methods like RSA to test whether brain regions represent task states similarly. If the main hypothesis is that neural activity reflects the (initial) learning rate itself, then a univariate analysis approach would seem more natural.*

We agree that task statistics are not the same as differences in learning rates. However, we do not consider this as a confound: The point of the differences in task statistics is exactly to generate differences in learning rates. With our paradigm, we deliberately tried to dissociate

variations in learning rate that were induced by learned environmental differences versus local task statistics. We tried to make this dissociation more clear, especially when discussing the fMRI results.

> *(3) For the neuroimaging results in particular, the specificity of some of the results (e.g. ventral striatum showing an effect of prediction error only in the low noise condition in the second half of task experience, only on the first trial) is a bit surprising. Additional justification of or context for these results would be useful to help readers gauge how expected or surprising these findings are.*

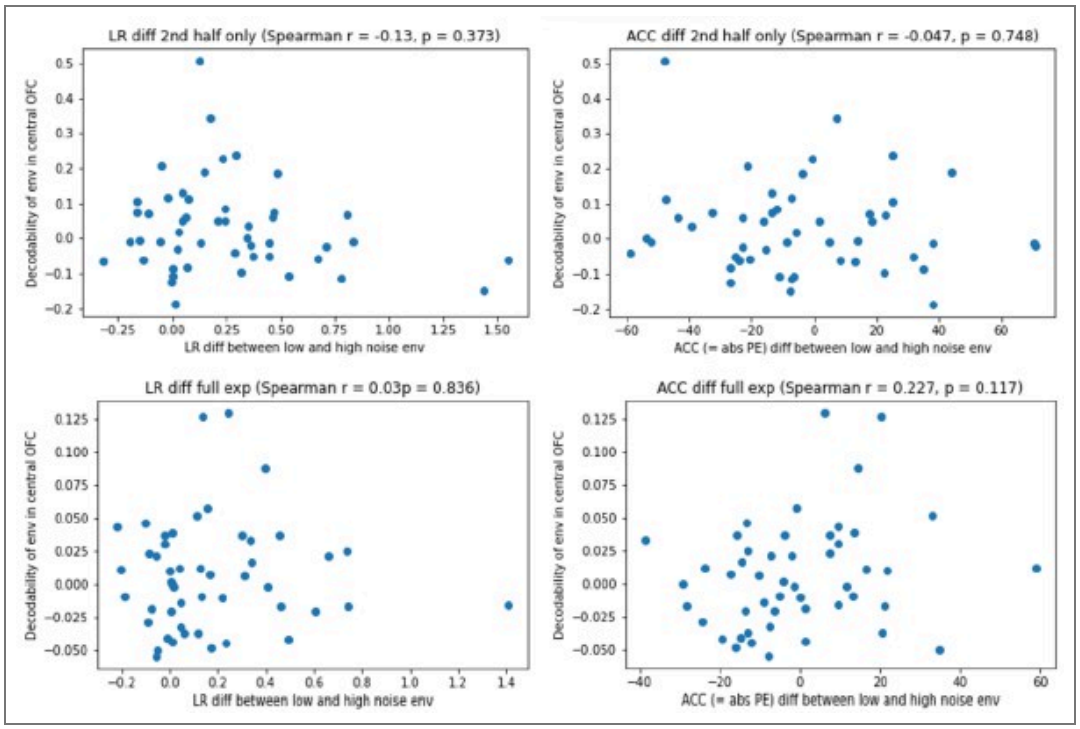
We agree some of these findings were unexpected. We now also highlight that while we expected the ventral striatum to be involved in prediction error processing, we had no strong a priori expectations regarding these further modulations by time and environment. We also tried to contextualize these interactions more.

> *(4) There are some methodological details that are unclear (e.g., how were the positions of the crabs selected relative to the location they emerged from? Looking at Figure 1C, it looks like the crabs spread out unevenly, and that the single position they emerge from is not necessarily at the center of the crab locations.) Additional detail and clarity would help address some unanswered questions (more details below).*

We clarified the experimental procedure at several places, and now added a video that helps illustrate the trial timeline better.

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations for the authors):***
>
> *(1) With regards to the primary weakness mentioned above, it would be nice to have some link between the brain signals of interest and upcoming behavior. For example, can you read something out of OFC that enables you to better predict what the participant will do next? Or even better, do so beyond any behavioral variability that is explained by the computational model?*

To link the behavioral with the fMRI data, we now correlated fMRI decoding accuracy with behavioral performance. We studied behavioral performance in two ways: the difference in high versus low noise environment learning rates, and mean accuracy (i.e., absolute prediction error). We correlated both measures with the decodability of the environment in the central OFC. Each correlation was calculated either in the full experiment, or only the second half. However, none of these correlations were significant (all $p > .1$; see plots in Author response image 1). Given the difficulty of interpreting this result, and our lack of statistical power for doing individual difference analyses, we decided not to report this analysis in the paper.



Author response image 1.

> *(2) A number of the learning analyses are based on splitting the session into halves. As a first pass, this seems like a reasonable thing to do, but I certainly wonder what the dynamics of the meta-learning actually look like, and it seems like the data collected would be sufficient to gain some insight into those dynamics through some sort of sliding window analysis.*

We thank the reviewer for this interesting suggestion, which was also raised by Reviewer 2. We now calculated the learning rate in a sliding window of 20 trials (i.e., trial x to x + 19), and provide revised figures for each experiment separately (Fig. 2E and Fig. 4E, respectively).

> *(3) The model selection procedures described make sense, but it would still be useful if the authors justified them by showing that they work in synthetic data (ie, generate a confusion matrix). I may be confused about what delta-SE is, but I'm confused about why two models with very different fits have the same value (211) for that metric.*

We report model recovery on synthetic data, which yielded model recovery rates of 100%, and added these to our Methods section. To clarify the Reviewer's second point, ΔSE is the standard error of the difference between a model's LOOIC and the top ranked model's LOOIC. There is no one-to-one mapping between the ΔSE and a model's LOOIC.

> *(4) Was the central OFC anatomical ROI overlapping with the cluster surviving in the whole brain analysis? I didn't see this mentioned in the text, and it certainly would be important for interpreting the two results together.*

The central OFC indeed overlapped with the cluster surviving whole brain analysis, which we report on page 17-18.

> *(5) The authors found regions that reflected learning rate at the "island presentation" phase of the task - it could be distinguishing this analysis and its meaning from other work that has focused on representations of learning rate at the time of feedback.*

We agree that this is an important distinction worth emphasizing. Therefore, we added the following lines to our discussion paragraph:

“Importantly, previous studies examined neural correlates of learning rates during outcome evaluation, where learning rates may be adjusted online as a function of locally experienced prediction errors (e.g., (Behrens et al., 2007; Browning et al., 2015; Nassar et al., 2012). In contrast, our RSA analysis targeted neural activity at island presentation, before any outcome information was available. At this moment, learning rates cannot be updated based on current feedback and instead reflected the retrieval of a previously learned, environment-specific learning-rate settings. This difference reflects our hypothesis that the OFC represents the latent states in a cognitive map of the task (Knudsen & Wallis, 2022; Moneta et al., 2024; Schuck et al., 2018; Wilson et al., 2014), which are expected to activate as soon as the agents can infer which task state it is in. Several studies have identified such “partially observable” task states in the medial OFC (Bradfield et al., 2015; Schuck et al., 2016; Tan et al., 2025; Wimmer & Büchel, 2019), in line with the region identified here (but see e.g., (Ongur & Price, 2000), for important anatomical distinctions between medial and lateral OFC and (Tan et al., 2025) for an example of related functions in lateral OFC). Our finding extends this notion by suggesting a link between OFC and meta learning, wherein meta-learned information becomes encapsulated in task states (Hattori et al., 2023; Moneta et al., 2024).”

> *(6) "Specifically, it showed a more negative response to larger (location) prediction errors, which is consistent with its documented role in showing a more positive response to more positive reward prediction errors (Calderon et al., 2021) - keeping in mind that being closer to the centre of where the crabs appeared (i.e., smaller location prediction errors) is less negatively or more positively surprising (i.e. smaller negative or larger positive reward prediction errors)."*
>
> *I found this sentence very hard to parse. Do PE responses in the high noise environment get "compressed" in their representation over time (ie, it takes a larger error to get the same BOLD response)? If so, this relates to claims made in Diederen 2016... but see also Mah 2024 Cell Reports, who fails to see learning rate encoded in DA system in striatum of rodents that appear to adjust their learning rates.*

Thank you for pointing to this. We agree that this sentence was hard to parse, and so we now split it in three revised sentences. We also agree with the Reviewer’s interpretation, and would like to thank the Reviewer for their useful literature suggestions which we now added to our discussion.

> *(7) Figure 7 should use a different color scheme because many of the activations just appear black, and I can't tell whether they are positive or negative. It was also notable in Figure 7A that regions are not visible, including ACC, which is typically thought to encode prediction errors in such paradigms. It would probably be useful for the authors to include a table of all clusters exceeding multiple comparisons correction and to on differences to other work examining absolute prediction errors. ACC does appear on the second trial, which made me wonder whether there were changes in the prediction error coding from first to subsequent trials.*

Thank you for pointing this out. We now revised our color scheme which we agree makes it much clearer now. Although the ACC is frequently implicated in prediction error–related signals (e.g., Behrens et al., 2007), models suggest that ACC responses more strongly reflect unsigned prediction errors, surprise, or the need for control and model updating (Alexander & Brown, 2019; Hayden et al., 2011; Silvetti et al., 2018). In our task, ACC activity only emerged on the second trial, when participants had formed an initial estimate and prediction errors could meaningfully signal the need to update internal models or control settings. We now

added a to the Discussion highlighting this distinction and relating our findings to this prior work emphasizing prediction errors and control-related signals in ACC.

> *(8) The authors suggest that fast learning would presumably occur in a neural activation space, whereas slow learning would occur through weight adjustments. This makes sense, but activity-based dynamics have been suggested to do rapid adjustments by encoding a "latent state" though (Razmi 2022 j neurosci) -- and such a latent state has been shown in OFC (Schuck etc)... but here OFC is more implicated in the slow learning. I am curious about whether authors could on this a bit in the discussion.*

Thank you for bringing up this interesting question. We can only speculate but a crucial factor is on which level of resolution tasks states operate. On the one hand “detailed” trial-level states are needed that map a specific sensory input onto a specific latent state and its value. Such states would change quickly, possibly through activation dynamics, and are in line with how they have been operationalized in Razmi or Schuck etc. On the other hand, successful task performance also needs “higher level” states that describe entire task phases or full tasks, as in the present experiment. Due to the different speeds of learning, it appears plausible that these would be learned with synaptic changes. We expand on this in the discussion as follows:

“Our finding extends this notion by suggesting a link between OFC and meta learning, wherein meta-learned information becomes encapsulated in task states (Hattori et al., 2023; Moneta et al., 2024). Consistently, OFC has been shown to represent task states (Moneta et al., 2024; Stalnaker et al., 2015; Wilson et al., 2014). While earlier evidence shows that the OFC represents concrete aspects of task states, such as task-relevant stimulus features (Schuck et al., 2016), we hypothesized that the OFC also represents more abstract aspects, such as learned, environment-specific learning rates. Indeed, we showed that the central OFC gradually came to represent these environment-specific learning rates (or the environment-specific statistics that drive them). While previous work speculated that these different levels could have different neural underpinnings (Sharpe et al., 2019), our findings indicate OFC might signal states on multiple levels. This does not imply identical learning dynamics; fast-changing trial-specific states might be learned through activity dynamics, while higher-level contextual states could involve synaptic plasticity.”

> *(1.9) Also, as a more minor point in the same section, the sentence about blocking synaptic plasticity in OFC sounded interesting, but should have a reference.*

Thank you for noticing, we now added the reference (Hattori et al., 2023).

> ***Reviewer #2 (Recommendations for the authors):***
>
> *(1) Additional links to prior literature: In terms of prior work in which there is something akin to more "global" adaptation, some examples of potentially relevant prior work include:*
>
> *McGuire, Nassar, Gold, & Kable (2014) Neuron*
>
> *D'Acremont & Bossaerts (2016) Cerebral Cortex*
>
> *Lee, Gold, & Kable (2020) Decision*
>
> *Bakst & McGuire (2021) JEP: General*
>
> *Bakst & McGuire (2023) Cognition*

We would like to thank the reviewer for pointing us to these different literature suggestions which we agree help us contextualize and discuss some of our findings better. We now refer to McGuire et al. (2014) when discussing the fMRI results, and d'Acremont & Bossaerts (2016)

when discussing potential alternative strategies in the high noise environment (the Reviewer's last point). Finally, we integrated the clearly relevant works of Bakst & McGuire (2021; 2023) and Lee et al. (2020) in our discussion of meta-learning different adaptive strategies.

> *(2) Individual differences: Though not always the focus of work on predictive inference, one common finding has been that there are pronounced individual differences in behavior (see, e.g., coefficients in Figure 2 in Nassar et al. 2019 eLife, or Figure 2 McGuire et al. 2014 Neuron, or Bakst & McGuire 2023 Cognition). There appears to be substantial variability between individuals in your data as well (i.e., Figure 2B, 4B, and the modeling figures). It would be interesting to see some direct exploration of this variability: baseline learning rate appears to differ between participants to a large extent, does their rate of adaptation (across trials within a block) also differ? Does their metalearning occur at different rates (in fact, do some participants not show evidence of appropriate meta-learning at all)?*
>
> *Relatedly, your computational modeling approach fits the six candidate models hierarchically, and therefore the reported results show the overall best fit for the group. It might be worthwhile to determine whether individuals have different best-fitting models. This could be another way to characterize the variability between individuals.*
>
> *In concert with this, it could be a useful complement to determine whether either the strength of the OFC neural similarity results or their time course reflects aspects of behavior. Put another way, is it the case that not only does OFC activity and behavior both come to reflect task structure, but that these changes happen to a similar extent and over a similar time course across individuals?*

We agree it would be highly interesting to investigate meaningful individual differences in both fast and slow adaptations in learning rate. However, our sample was not set up and is underpowered to conduct such analyses. In response to a similar by Reviewer 1, we did run correlational analyses between differences in learning rate, performance accuracy, and the responsiveness of the OFC. However, none of these analyses yielded a significant effect. We decided to not include these results in the paper, for reasons of statistical power, but we report them in Author response image 1.

> *(3) fMRI:*
>
> *(3a) The primary finding in OFC is restricted to the central OFC. The manuscript would benefit from additional explanation regarding this specific subregion.*

Thank you for bringing up this important distinction. In the discussion we now clarify as follows:

"This difference reflects our hypothesis that the OFC represents the latent states in a cognitive map of the task (Wilson et al., 2014; Schuck et al. 2018; Knudsen & Wallis, 2022; Moneta et al, 2023), which are expected to activate as soon as the agents can infer which task state it is in. Several studies have identified such "partially observable" task states in the medial OFC (Schuck et al., 2016; Bradfield et al., 2015; Wimmer et al., 2019; Tan et al., 2025), in line with the region identified here (but see e.g., Öngur & Price, 2000, for important anatomical distinctions between medial and lateral OFC and Tan et al., 2025, for an example of related functions in lateral OFC)."

> *(3b) Though the main clusters visible in Figure 6 are the occipital and OFC clusters, there appear to be others. Did other clusters indeed rise to statistical significance in the whole-brain analysis? If so, is there a reason they aren't included or discussed?*

All clusters visible in Figure 6C survived FDR correction. However, we refrained from interpreting these other clusters, because we had no prior hypotheses about them like we did for the OFC.

> *(3c) Why do you posit that the ventral striatum becomes less sensitive to RPE on the second trial over time? And why is the ventral striatum only sensitive to RPE in the low noise environment generally?*

We reasoned the ventral striatum should be more responsive to more positive reward prediction errors. While we further assumed this response could be modulated by both time and environment, we would like to emphasize that we had no specific hypotheses about the direction of this modulation. We now also make this clearer in the manuscript. This being said, we believe both the pattern that its responsiveness to the second trial decreases over time, and the pattern that it was most sensitive to the low noise environment, can be considered fitting with its broader involvement in coding behaviorally relevant reward prediction errors. Namely:

First, we believe that as the participants learn more about the global reward structure of the task, they should obtain a better understanding of the fact that, per round, all crabs always center around a fixed mean. Therefore, the first RPE is most behaviorally relevant, and every later RPE has an exponentially decreasing relevance. As participants obtain more experience with this aspect of the task over time, the VS should show a lower responsiveness to the second RPE over time.

Second, as participants learn more about the local differences between the three different environments, they should learn that especially in the low noise environment, RPEs are most behaviorally informative. That is, in this environment it makes most sense to have a high learning rate and thus let the RPEs substantially inform the placement of the cage on the next trial. Accordingly, participants showed that the ventral striatum was most responsive to RPEs in these environments.

> *(4) Methods*
>
> *(4a) This section could generally benefit from some proofreading.*

We now proofread the method section.

> *(4b) The main results text states that 49 participants performed Experiment 1, while the methods section reports 50 participants. Which is correct?*
>
> *(4c) Following this, on page 8, statistical results are reported with a df = 49 (which would be appropriate only if n=50).*

The correct sample size was actually 50, we adjusted the text and degrees of freedom where incorrect accordingly (note: only text is in track changes, but degrees of freedom were also changed accordingly).

> *(4d) Additionally, I am a bit surprised by the Experiment 1 findings that learning rates on the second trial were significantly different between low and high noise conditions, in that the effect size found using all trials was stronger than both the first half of trials (no significant effect) and the second half (significant but weaker than all trials). Are these all the same type of statistical test? Double-checking the statistics might be worthwhile.*

It is not the effect size that is larger across the full experiment, but the t-statistic. This is possible because a t-statistic depends on both effect size and noise estimate, and the latter is smaller with more data.

> *(4e) The methods and results both state that the five crabs always emerged from one position in the sand. How were the locations of the crabs selected relative to this position? Looking at Figure 1C, it looks like the crabs spread out unevenly, and that the single position they emerge from is not necessarily at the center of the crab locations.*

The crabs did indeed spread out evenly. However, we can see how the graphic in Figure 1C can be confusing, as two crabs are shown to be caught, which breaks the symmetry of the dispersion (because some crabs can run away after the even spreading phase, see Methods). We emphasized the even spreading more clearly in the new version of the paper. We think the flow of events will be much clearer with our newly added animation (Video 1).

> *(4f) The methods section states that the crabs "spread out to cover the same proportion of the screen width as the cage (18.75%)" (page 23). The corresponding visual in Figure 1C appears to show something different.*

This looks different because the graphic illustrates the last 500 msec, where crabs can run away (see also response to 4e, and the novel animation that was added).

> *(4g) Information on the timing of the trials would be useful to include in Figure 1C or similar.*

The reader can find this information in the Methods section. We chose not to include it in the caption to avoid information overload.

> *(4h) The methods section specifies that there was a 3-7s ITI after the first and second trials of each block. How was the ITI selected for each trial? Were there ITIs between the other trials? If so, what were they?*

The ITIs were selected from a truncated exponential distribution. This selection was not random, but rather a distribution was carefully constructed for each environment (and event of interest: boat presentation, first trial of each block, second trial of each block) separately to ensure that enough longer ITIs were selected for each environment (and event of interest). Of course, the order in which the ITIs were used across blocks, was random. The same approach was used to determine the duration of the presentation of the boat at the start of each block. There were no ITIs after later trials.

> *(4i) Please provide a link to the data and analysis materials on OSF in the text.*

We now provide a link to the data and analysis materials in our methods section.

> *(4j) In the methods section, there are some references to information provided "below" (page 26: "The two approaches resulted in different posterior densities (see below) for estimate uncertainties, but in similar posterior densities (see below) for learning rates..."). Where in the paper is this referencing?*

We indeed did not detail this further as we considered it not further relevant to our main study, and now removed the references to “below”.

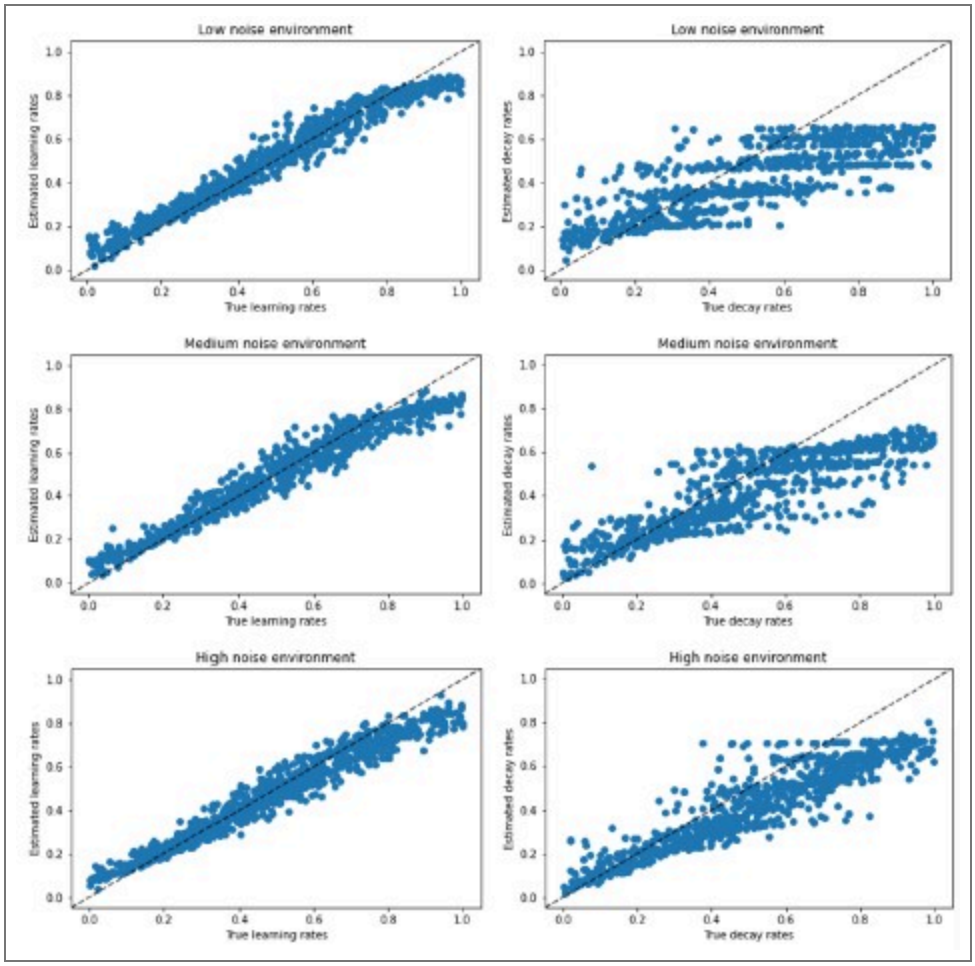
> *(4k) The methods section specifies using uniform priors between the lower and upper bounds of the relevant parameters. This seems likely to be 0 and 1, but should be listed explicitly.*

Thank you for noticing. We now added this to our manuscript.

> *(4l) For parameter recovery, correlations are provided to indicate effective recovery. These correlations are indeed high and suggest excellent recovery, but correlations wouldn't reveal if there was systematic over- or underestimation occurring. It might be*

> *useful to provide some visualizations of the parameters and their estimates to speak to this potential issue.*

We now visualize the parameter recovery results in Author response image 2, which show that, indeed, there was a slight underestimation of the decay rates, but not the learning rates. Importantly, our main analyses and results all pertain to the learning rates, and we never made hypotheses or conclusions about the decay rates.



Author response image 2.

> *(4m) The methods section ends with a reference to a reward localizer (page 32). This localizer doesn't appear to be mentioned/used elsewhere.*

Indeed. We implemented the localizer because we wanted to independently identify reward processing areas. However, this localizer did not succeed in localizing a reward area (no significant results), possibly due to the fact that (1) it was performed by the end of the experiment when participants may have been fatigued, and (2) there was no learning component in this localizer task. For these reasons, we did not use it after all.

> *(5) Analysis:*

> *(5a) Did you consider fitting a Bai model that only allowed for environment-specific initial learning rates (with a non-environment-specific decay rate)? Given that the data (e.g., Figure 2, Figure 4) seems to support differences in initial learning rate but not necessarily a difference in the rate of change, it might be worthwhile to see whether a model like that fits best.*

We now fitted this extra model, which we called the semi-environment-specific Bai model. See Author response tables 1 and 2 for result in experiments 1 and 2, respectively) for the results. This new model has the best (in Experiment 2) and second-to-best (in Experiment 1) LOOIC. In a way, this is not surprising, because the model formulation is entirely based on the data. We think that we can draw the same substantive conclusions with or without this extra model, so for simplicity we did not include this new model in the paper itself.

Model	LOOIC	SE	ΔLOOIC	ΔSE
Environment-specific Bai model	28735	404	0	0
Non-environment-specific Bai model	28582	441	153	65
Environment-specific Rescorla-Wagner model	28436	398	299	67
Non-environment-specific Rescorla-Wagner model	28172	440	563	112
Environment-specific Kalman filter	27853	493	882	211
Non-environment-specific Kalman filter	27698	521	1037	211

Author response table 1.

Note. Models are ranked in descending order according to how well they fit the data. LOOIC refers to a model's approximated expected log pointwise predictive density. Higher values indicate higher out-of-sample predictive fit. SE refers to the standard error of a model's LOOIC. ΔLOOIC refers to the difference between a model's LOOIC and the top ranked model's LOOIC. ΔSE refers to the standard error of the difference between a model's LOOIC and the top ranked model's LOOIC.

Model	LOOIC	SE	ΔLOOIC	ΔSE
Environment-specific Bai model	34725	142	0	0
Non-environment-specific Bai model	34682	158	43	34
Environment-specific Kalman filter	34592	160	133	42
Environment-specific Rescorla-Wagner model	34567	136	158	32
Non-environment-specific Kalman filter	34523	174	202	49
Non-environment-specific Rescorla-Wagner model	34434	154	291	51

Author response table 2.

Note. Models are ranked in descending order according to how well they fit the data. LOOIC refers to a model's approximated expected log pointwise predictive density. Higher values indicate higher out-of-sample predictive fit. SE refers to the standard error of a model's LOOIC. ΔLOOIC refers to the difference between a model's LOOIC and the top ranked model's LOOIC. ΔSE refers to the standard error of the difference between a model's LOOIC and the top ranked model's LOOIC.

> *(5b) If part of the goal is to investigate whether there is a distinct local change in LR between conditions (dependent on prediction errors), then there might be more direct*

> *ways of doing so as a complement to the modeling approach. One potential way could be to visualize the LR or change in LR as a function of PE.*

We agree that it's beneficial to use a direct (model-free) approach to represent learning rate as a function of condition; that is also part of our approach. For example, see Figures 2, 4, which shows learning rate as a function of condition, but in a model-free manner. We think learning rate as a function of prediction error is less informative, because the idea is that prediction error can (in Kalman-filter terminology) be indicative of either noise variance or process variance, and participants are able to distinguish between them. This is also why we constructed the conditions in such a way that on the very first trial, prediction errors were on average the same across conditions. The fact that participants did respond appropriately to prediction errors on the very first trial (i.e., larger updates or learning rates in the low noise condition), suggested they are able to assign the prediction error to process variance (in the low noise condition) versus noise variance (in the high noise condition).

> *(5c) In addition to looking at the evolution of LR across trials within a block separated by task epoch (i.e., Figure 2C-D & Figure 4C-F), the structure of the task would lend itself very nicely to visualizing the evolution of the second trial LR on its own across instances. This could provide additional insight into the meta-learning process.*

We thank the reviewer for this interesting suggestion, which was also raised by Reviewer 1. We now calculated the learning rate in a sliding window of 20 trials (i.e., trial x to x + 19), and provide revised figures for each experiment separately (Fig. 2 and 4, respectively).

> *(6) The environment-specific Bai model appeared to become less good at capturing participant behavior with increased environmental noise. Why do you think this is?*

We thank the reviewer for raising this point. In this environment, individual outcomes are considerably less indicative of the latent mean, which may reduce the usefulness of the trial-by-trial, prediction-error–driven learning-rate adjustments that we see in the other environments. Under such extreme conditions of variability, people may rely less on delta-rule updating and more on alternative strategies (D'Acremont & Bossaerts, 2016; Reynders et al., 2026), such as exploratory adjustments or heuristics that are not explicitly captured by the Bai model but also outside the scope of the present paper.

https://doi.org/10.7554/eLife.108223.2.sa0