A nonlinear relationship between prediction errors and learning rates in human reinforcement-learning

Boluwatife Ikwunne[1], Jolie Parham[1], Erdem Pulcu[1,2]

[1]University of Oxford

Department of Psychiatry

[1]Psychopharmacology and Emotion Research Lab

[2]Computational Psychiatry Lab

Correspondence: Dr. Erdem Pulcu

email address: erdem.pulcu@psych.ox.ac.uk; pulerd@gmail.com

Department of Psychiatry

Neurosciences Building

Warneford Hospital

Oxford OX3 7JX

United Kingdom

Abstract:

Reinforcement-learning (RL) models have been pivotal to our understanding of how agents perform learning-based adaptions in dynamically changing environments. However, the exact nature of the relationship (e.g. linear, logarithmic etc.) between key components of RL models such as prediction errors (PEs; the difference between the agent's expectation and the actual outcome) and learning rates (LRs; a coefficient used by agents to update their beliefs about the environment) has not been studied in detail. Here, across (i) simulations, (ii) reanalyses of readily available datasets and (iii) a novel experiment, we demonstrate that the relationship between PEs and LRs is (i) nonlinear over the PE/LR space, and (ii) it can be better accounted for by an exponential-logarithmic function that can transform the magnitude of PEs instantaneously to LRs. In line with the temporal predictions of this model, we show that physiological correlates of LRs accumulate while learners observe the outcome of their choices and update their beliefs about the environment.

Keywords: reinforcement-learning, uncertainty, LRs, prediction errors, computational modelling

Introduction

We are living in a dynamic world in which probabilistic relationships between cues and outcome events create different sources of uncertainty[1] (e.g. not all grey clouds bring rain). Living in an uncertain world continuously probes learning systems in the brain, guiding agents to make better decisions. This is a type of value-based decision-making is critical for survival in the wild and long-term evolutionary fitness. Consequently, reinforcement-learning (RL) models describing cognitive/computational processes underlying learning-based adaptations have been pivotal in behavioural[2,3] and neural sciences[4-8], as well as machine learning[9,10].

A number of previous studies showed that human learners adjust their learning rates (LRs) based on the information content of outcomes[11,12], utilising a higher learning rate for outcomes more informative for predicting the future state of the environment. In computational neuroscience, LRs are numerically estimated from participant choices by using different RL algorithms[13-15] —which are in effect approximations of the Bayesian models[16,17] that can also estimate otherwise hidden generative rules of dynamically changing environments. The Rescorla-Wagner (RW) algorithm appears to be the most widely used RL model[13] in the literature:

$$\tilde{p}_{(t+1)} = \tilde{p}_{(t)} + \eta(\Theta_{(t)} - \tilde{p}_{(t)}) \quad (1)$$

where $\tilde{p}$ is the agent's probability estimate for an outcome, Θ is the actual outcome (e.g. win vs null), η is the learning rate and *t* is the trial number. However, a key limitation of the RW model is that it is commonly fitted to participant choices within a block of an experimental task, estimating participants' average LRs during that period. This is an approach that uses the RW model as a measurement tool, implemented by a number of key RL studies[4,11,12,16] consistently showing that human LRs are higher in volatile (i.e. environments in which contingencies change), relative to stable environments in which contingencies stay the same. However, these findings also indicate that LRs are not constant. As a result, approximating a single learning rate from a block of an experiment may not be the most ideal approach for understanding mechanisms underlying learning-based behavioural adaptations. The family of hybrid Pearce-Hall models[14,15,18] aim to mitigate this limitation by updating LRs on a trial-by-trial basis. For example, one common implementation is based on a weighted integration of the LRs from previous trials with the magnitude of the absolute value of the prediction errors (PEs, i.e. the difference between the agent's expectation and the real outcome):

$$\eta_{(t+1)} = \omega\eta_{(t)} + (1-\omega)\,|\,(\Theta_{(t)} - \tilde{p}_{(t)})\,| \tag{2}$$

where ω is a free parameter estimated between 0 and 1 and determines how much weight is given to the learning rate (η) on the current trial and the absolute value of the PE (i.e. the unsigned PE) to update the learning rate into the next trial. Here, it is important to highlight that in the majority of the RL models, the absolute value of the PEs ($|(\Theta_{(t)} - \tilde{p}_{(t)})|$) is commonly used as a proxy for the agent's surprise. The model would then update the agent's estimates of the reward probability into the next trial:

$$\tilde{p}_{(t+1)} = \tilde{p}_{(t)} + \lambda\eta_{(t)}(\Theta_{(t)} - \tilde{p}_{(t)}) \tag{3}$$

where λ is another free-parameter commonly used for smoothing the transition of the probability update. While both RW and Pearce-Hall models are commonly used in the literature, these were not developed ground-up considering how the relationship between prediction errors and LRs should be (e.g. linear, logarithmic or exponential, etc.).

Learning-based behavioural adaptations in response to different sources of uncertainty

Existing literature suggests that different sources of uncertainty dictate which style of learning behaviour would be optimal in a given environment[1,19,20]. *Expected uncertainty* relates to intrinsic variability in the occurrence or the nature of an event over and above what could be explained by any learned association. For example, if an outcome follows a stimulus on only 30% of occasions, or if the magnitude of the outcome is inherently variable (e.g. day-to-day changes in air temperature)[1], it would be difficult to accurately predict the frequency or the magnitude of the outcome on any given occasion. In other words, the higher the expected uncertainty the less informative each particular event is; since expected uncertainty erodes the information content of individual events[1]. In environments dominated solely by expected uncertainty, the optimal learning behaviour would be to estimate cue-outcome associations over a long period of time utilising a relatively lower learning rate. However, there can be instances when learned associations change over time. In these cases, the true association between cues and outcomes is more difficult to estimate because it is not stable. This is commonly known as *unexpected uncertainty* (or "*volatility*"[16]). Unexpected uncertainty effectively reduces how informative the previous events are during learning. In other words, previous events become increasingly less informative the higher the unexpected uncertainty in the environment. The optimal behavioural adaptation to this type of uncertainty is to put more weight on recent events, as events further back in history are likely to be less informative about the future state of the environment. Particularly in environments with high unexpected and low expected

uncertainty, tuning LRs to the prediction errors (PEs) would be an efficient way of interacting with the environment. This is because during periods of low expected uncertainty (i.e. if the environment is not noisy), the PEs would be relatively small, and agents can discount them by utilising a lower learning rate. However, when the PEs are reasonably high, this would signal that there must be a dramatic change happening in the underlying structure of the environment. In these situations, agents can quickly adapt by utilising a higher learning rate. Although these conceptual descriptions indicate a relationship between PEs and LRs, the exact nature of this relationship is yet to be explored in detail. Below, we evaluate the suitability of novel update rules for RL, based on a nonlinear relationship between LRs and the magnitude of the PEs.

A parabolic relationship between learning rates and prediction errors

In RL tasks with probabilistic contingencies (i.e. reward versus no reward), the PEs are always between -1 and 1. For example, if the environment returns no rewards (i.e. a vector of null outcomes: [0, 0, 0, 0 ...]) yet the agent is in absolute denial, believing that s/he is in a highly rewarding environment, the lowest value that the PEs can be is a vector of -1s. Another widely established assumption is that there should be no additional learning when the PE is zero. That is if the agent can consistently predict reward contingencies with 100% accuracy, then there is nothing new to be learned from the environment and the effect of a constant learning rate should diminish (Eq.1). Also considering the nonlinearity in wide majority of cognitive processes (e.g. temporal discounting[21,22], risk perception[23,24], stochastic choice[25,26] etc.) in human and non-human primates, a simple functional form satisfying all of these assumptions proposes a parabolic relationship between the PEs and LRs:

$$\eta = (\Theta_{(t)} - \tilde{p}_{(t)})^2 / 4\kappa \tag{4}$$

where (0,κ) is the focal point of the parabolic curve on the *xy* plane, adjusting how steeply PEs transform to LRs (Figure 1), and it is the free parameter to be estimated from participant choices.

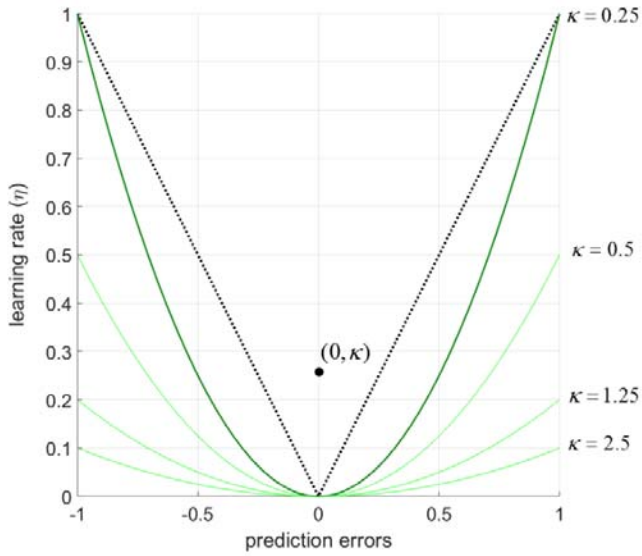


Figure 1. A parabolic relationship between prediction errors and learning rates. In environments with high unexpected but low expected uncertainty, one would assume a somewhat linear relationship between prediction errors and LRs (in a V shape designated by the black dashed lines), meaning that agents should update their beliefs proportionately to how far away their beliefs about the environment was relative to the reality. The proposed algorithm approximates this relationship under the control of a single free-parameter (thick green line, κ=0.25). The free-parameter κ also determines the focal point of the parabolic curve (i.e. the black dot at (0, 0.25) on the *xy* plane is the focal point for the thick green parabolic curve). Increasing values of κ flattening the relationship between prediction errors and LRs, leading to lower LRs at a given magnitude of the prediction error. A few examples are illustrated by light green lines at κ=0.5; 1.25 and 2.5. Parameter values smaller than 0.25 could allow prediction errors to be converted to LRs even more steeply.

Then the update algorithm (i.e. updating PEs with corresponding LRs on each trial) can be expressed as:

$$\tilde{p}_{(t+1)} = \tilde{p}_{(t)} + (\Theta_{(t)} - \tilde{p}_{(t)})^3 / 4\kappa \tag{5}$$

We will refer to this model as the cubic model in the rest of the manuscript. Here, the free-parameter κ accounts for individual variability in how steeply LRs increase at a given magnitude of PE. This cubic update rule would allow agents to adjust their behaviour dynamically as soon as unexpected changes happen in the environment (i.e. those leading to higher PEs). However, in noisy environments (i.e. those with high expected uncertainty) scaling LRs to the PEs would not be an ideal way to learn from the environment. This is because, in stable environments with high expected uncertainty, agents need to discount PEs as these would be noisy and not informative of the future state of the environment. The model can also account for this behaviour at higher values of κ (Figure 1). From this perspective, the parameter κ is also under the influence of the agent's global inference about the level of expected uncertainty (i.e. noise or the standard deviation (σ) of the generative process) in the environment[19]. A recent analysis of normative learning models demonstrated that LRs decay exponentially in noisy environments with high expected uncertainty[27]. Therefore, it is also

plausible to assume that the higher-level relationship between expected uncertainty (i.e. environmental noise (σ)) and the free-parameter κ can be in an exponential form:

$$\kappa = f(\sigma) = xe^{\beta\sigma} + \varepsilon \tag{6}$$

which would allow the model to tune LRs with respect to the magnitude of prediction errors at a given level of environmental noise. Under these assumptions, the nonlinearity of LRs guiding behavioural adaptations could resemble Figure 2, illustrated below. The distribution of these LRs is globally in line with results reported by other research groups which showed a linear-stochastic relationship between LRs estimated from participant choices and prediction errors as a function of expected uncertainty/environmental noise[28].

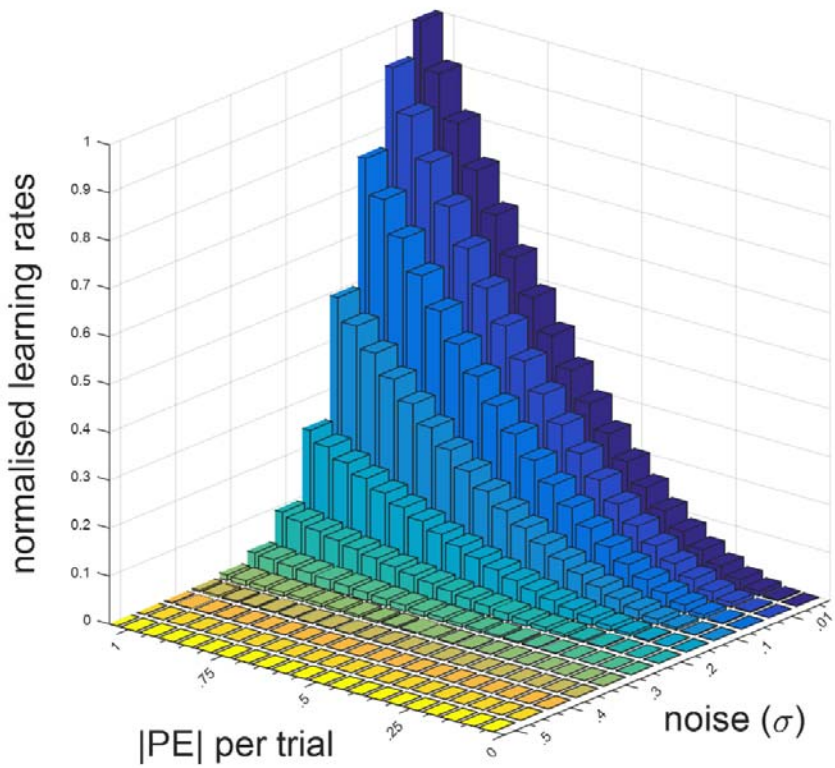


Figure 2. Nonlinearity in trial-wise LRs as a function of prediction errors and environmental noise. Increasing levels of expected uncertainty (σ) diminish the relationship between PEs and LRs. In these situations, the agents should no longer attribute high magnitude prediction errors to indicate informative changes in the environment but should simply treat them as noisy fluctuations which are not informative. These are in line with normative learning models which show that LRs decay exponentially as expected uncertainty in the environment increases. In noisy environments in which the rewards are delivered in the normalised space (i.e. between 0 and 1), maximum entropy is reached if the mean reward rate is 0.5±σ. One viable function, perhaps out of many possible solutions, accounting for the relationship between κ and σ is κ=$f(\sigma) = \frac{\pi e^{7\pi\sigma}}{10^3} + \frac{\gamma}{6}$. In learning tasks in which the expected uncertainty is experimentally manipulated, it may be conceptually more accurate to estimate σ as a free-parameter from participant choices rather than κ. The figure demonstrates that the σ parameter can be estimated reliably between .01 and .4, covering the whole learning rate space. Note that for outcome magnitudes within the normalised space (i.e. [0 1]), values of σ higher than .3 would generate outcomes outside these boundaries. γ denotes the Euler-Mascheroni constant.

A data driven approach for evaluating the nature of nonlinearity between LRs and prediction errors

Another approach with which one can extract trial-wise LRs is by exploiting the mathematical properties of a 2-parameter [probability] weighting function[29]. This approach would assume an exponential-logarithmic relationship between the unsigned PEs and LRs and allows even more flexibility than the parabolic relationship described by the cubic model above, as it is free of any *a priori* assumptions about the exact shape of the relationship between PEs and LRs:

$$\eta = e^{(-\delta(-\log(|\Theta_{(t)} - \tilde{p}_{(t)}|))^{\lambda})} \tag{7}$$

and the rest of the update rule remains identical to the RW model (Eq.1). The parameters δ and λ are to be estimated from participant choices (both parameters can be estimated within boundaries [0, ∞], Figure 3). This model, which can cover the whole PE-learning rate space with different nonlinear trajectories (Figure 3), can help with further scrutinising whether the relationship between PEs and LRs is indeed parabolic (Eq. 5). The strength of this approach is that it would allow greater flexibility to account for different trajectories, and it can be fully reduced to the RW model as the model is able to generate constant LRs over the prediction errors, as shown by horizontal lines in Figure 3.

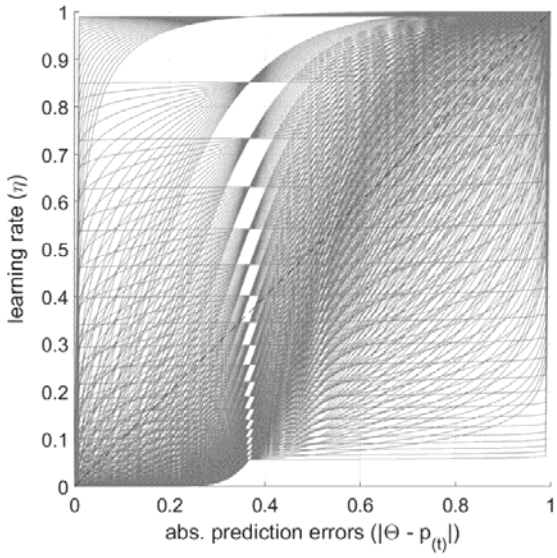


Figure 3. Nonlinearity in trial-wise LRs as a function of prediction errors under an exponential-logarithmic assumption. An exponential-logarithmic weighting function allows greater flexibility in capturing individual variability in how PEs shape LRs. The parameters δ and λ are coarsely sampled for illustration purposes. When the parameter combinations produce horizontal lines (e.g. y=0.2) over the PEs (i.e. the x-axis), the model is fully reduced to the Rescorla-Wagner rule with a constant learning rate.

Results

Simulated data. We simulated the behaviour of RL models (i.e. Rescorla-Wagner, hybrid Pearce-Hall, cubic and exponential-logarithmic models) in a task environment in which unexpected (i.e. volatility) and expected (i.e. noise) uncertainties were manipulated independently in different segments of the task environment (2x2 factorial design: high volatility low noise, low volatility high noise covering all possibilities, Figure 4). LRs were estimated under each RL model by minimising the average PE per trial throughout the task environment (t=1500 trials). In our initial analysis, we performed a procedure for minimising prediction errors based on the full trajectory of outcomes (including the future state of the environment) rather than adapting learning rates based on local/ past information, as at this stage we did not model agents making decisions under different models and arbitrarily assigned learning rates. It is worthwhile to highlight that our approach tests the upper boundary of how well simulated agents could perform in dynamically changing environments under each model which could be understood as model flexibility, which is an important metric for describing adaptive behaviour.

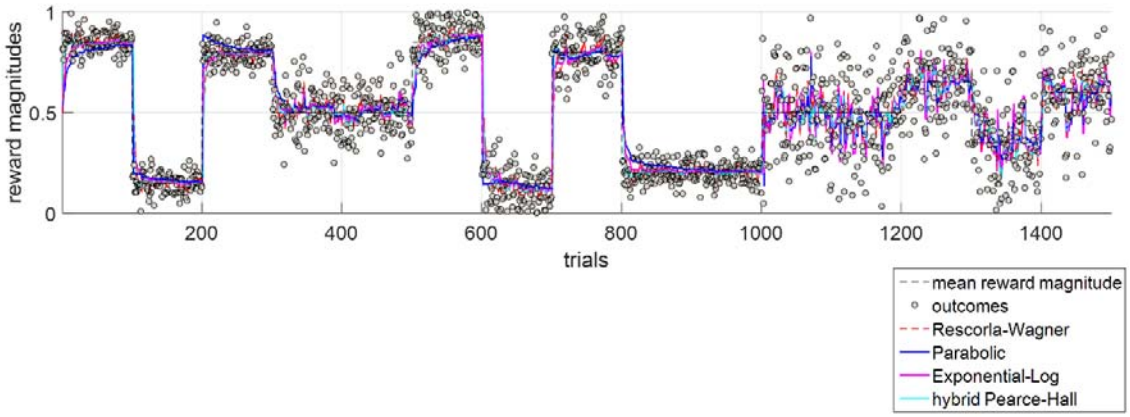


Figure 4. The behaviour of different reinforcement-learning models in a task environment in which unexpected and expected uncertainties were independently manipulated. On each trial, the task environment returns a reward outcome between 0 and 1 (y-axis). The outcomes were generated to be variable from one trial to the next. The first 300 trials are volatile (i.e. the mean of the underlying Gaussian distribution jumps to a different point in every 100 trials) with low noise (SD=0.05). The subsequent 200 trials are a low volatility environment (i.e. the mean of the Gaussian distribution generating the rewards stay the same) with a higher level of noise (SD=0.1). After 500 trials, the environment becomes volatile again, this time with high noise level (SD=0.10). The next 200 trials are a low volatility environment with low noise (SD=0.05). In the last 500 trials the environment is noisier (SD>=.15), which further obscures the transition between stable and volatile periods. All models converge reasonably well with the actual mean of variable rewards. The learning rate for the Rescorla-Wagner model (η, Eq.1) is 0.32. For the hybrid Pearce-Hall model, ω (Eq.2) is 0.48 and λ is (Eq.3) is 1.56. For the cubic model κ is 0.11 (Eqs. 4-5). For the exponential-logarithmic model, the parameters δ and λ are 0.83 and 1.45, respectively (Eq.7).

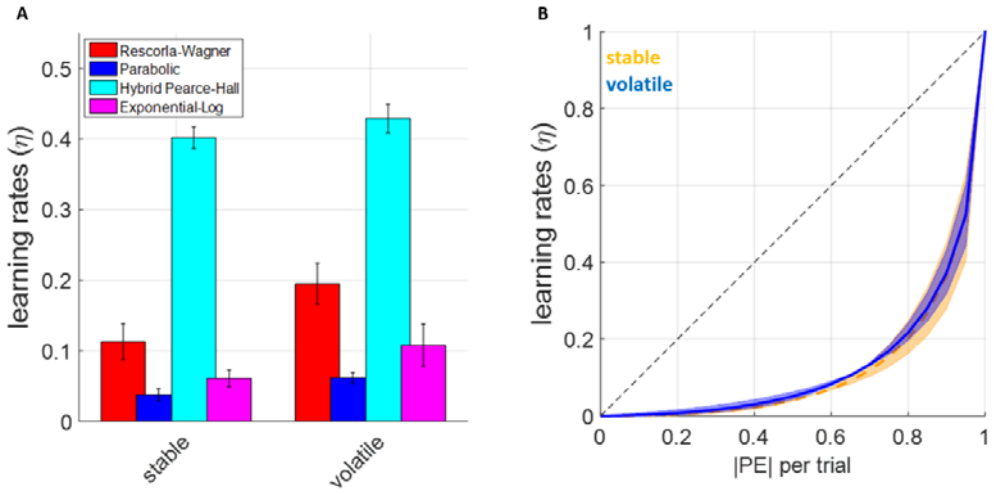
In the simulated task environment, all models converge reasonably well with the true average reward rate (Figure 4). The magnitudes of prediction errors under each model were not statistically different ($F(4, 7495) = 0.32$, $p = 0.863$, Supplementary Figure 1), meaning that all models can account

249 for behavioural adjustments comparably well. The nonlinear relationship between PEs and LRs under
250 the cubic and the exponential-logarithmic models are illustrated in Supplementary Figure 2.

251 The hybrid Pearce-Hall, the cubic and the exponential-logarithmic models can estimate LRs on trial-
252 by-trial basis. The PEs and LRs estimated under each model (based on model behaviours shown in
253 Figure 4) were decorrelated ($-.01 < r\ (1498) < .01$, all $p > .80$). However, LRs estimated under the hybrid
254 Pearce-Hall model (mean±SD = 0.096 ± 0.071) were significantly higher than the ones estimated
255 under the cubic (mean±SD = 0.042 ± 0.094) and exponential-logarithmic models (mean±SD = 0.076 ±
256 0.12, $F\ (2, 4497) = 119.07$, $p < .001$). Nevertheless, these learning rate estimates from different
257 models were very significantly correlated (all $r\ (1498) > .68$, $p < .0001$). Taken together, these results
258 indicate that 1-parameter cubic and 2-parameter exponential-logarithmic models can be viable
259 alternatives to the existing RL models. Here, it is worthwhile to highlight that one of the main
260 differences between these models is how quickly they would adjust their behaviours in response to
261 environmental change. Under the cubic and exponential-logarithmic models, LRs accelerate as soon
262 as there is a dramatic environmental change, therefore leading to large PEs, and LRs subsequently
263 decay as the magnitude of PEs get smaller. Comparatively, under the hybrid Pearce-Hall model
264 acceleration and deceleration of LRs happen relatively more gradually. This is because LRs here are
265 weighted integration of LRs and the PEs from the previous trials (Eq. 3, Supplementary Figure 3), as
266 determined by a weighting constant (Eq.2) that is estimated from participant choices. Significantly,
267 although each of these models describe an efficient way of interacting with the environment (i.e.
268 accelerating LRs only in cases in which the changes in the environment are informative), notably,
269 they make different predictions about the temporal properties of LRs (Supplementary Figure 3). In
270 the hybrid Pearce-Hall model the information is fed forward as the learning rate is a secondary
271 process updated in parallel (Eqs. 2-3) in addition to estimated reward rate, whereas cubic and
272 exponential-logarithmic models assume that prediction errors would trigger LRs instantaneously
273 during the feedback period of the current trial. We tested this prediction on temporal properties of
274 LRs subsequently in a novel RL experiment with pupillometry.

275 **Human behaviour.** First, we investigated how well nonlinear RL models explain human behaviour in
276 two previously published data sets. The first dataset contained healthy volunteers (N=18) who
277 completed a RL task with stable (120 trials) or volatile (170 trials) reward contingencies in different
278 task blocks[16]. In the original work by Behrens and colleagues (2007), participants were asked to
279 choose between two fractals which returned variable reward amounts on each trial or a null
280 outcome, based on the underlying reward probabilities that participants needed to learn. In the
281 learning task, the reward magnitudes associated with each fractal were generated randomly. In the
282 current reanalysis of this data, the stochastic choice function fitted to participant choice behaviour

283 was identical across all RL models. More specifically, participants' decision values were computed by
284 integrating their reward probability estimates (computed under each RL model) with their
285 corresponding reward magnitudes associated with each option (modulated by a single power utility
286 parameter accounting for participants' risk attitude[23]) and calculating the expected value difference
287 between two options and choosing between them stochastically following the choice probabilities
288 generated by a sigmoid function[25]. This is a very common approach in modelling participant choice
289 behaviour in this task.



290

291 **Figure 5. The LRs estimated under each reinforcement-learning model. (A)** Average LRs estimated under each model.
292 Hybrid Pearce-Hall model estimates higher LRs relative to the Rescorla-Wagner model, whereas cubic and exponential-
293 logarithmic models estimate overall lower LRs. **(B)** The relationship between PEs and LRs estimated by the exponential-
294 logarithmic model reveals a nonlinear trajectory very similar to what is proposed by the cubic model. Volatility of different
295 task blocks did not seem to influence learning rate trajectories.

296 Overall, all RL models estimated higher LRs in the volatile relative to the stable block, replicating the
297 findings previously reported by Behrens and colleagues (2007). Analysis of model log likelihoods by
298 fitting a 2x4 repeated measures analysis of variance (ANOVA) did not reveal any significant main
299 effect of block type (i.e. volatile or stable; $F(1,68)=.014$, $p=.907$) or model ($F(1,68)=.018$, $p=.997$), or
300 any significant block type x model interaction ($F(3,68)=.002$, $p>.999$). These results suggest that all
301 models tested here fit both volatile and stable environments comparably well. However, the
302 difference in LRs between stable and volatile blocks were only statistically significant under the RW
303 and cubic models ($t(34)$ values 2.114 and 2.079, p values 0.042 and 0.045, respectively). The average
304 LRs from the hybrid Pearce-Hall and the exponential-logarithmic models did not capture a
305 statistically significant volatility effect ($t(34)=1.057$ and 1.460, $p=.298$ and .153, respectively).
306 Similar to the simulated results reported in the preceding section, the hybrid Pearce-Hall model
307 generated much higher learning rate estimates relative to other models (Figure 5A).

308

Exponential-logarithmic model provides greater flexibility capturing behavioural adaptations

In order to tease apart the differences between the cubic and the exponential-logarithmic models further, we analysed another set of previously published data[30]. In this experiment, Vaghi and colleagues (2017) reported results from 25 healthy volunteers who tried to catch a particle initiating from the centre of the circular grid by correctly positioning a virtual bucket in order to collect as much points as possible (i.e."the bucket task", Figure 6A). This experimental design is rather similar to the simulated experiment shown in Figure 4. Participants completed 300 trials in total across 4 blocks of 75 trials, which meant that overall the environment changed more frequently than our simulated environment. Although the trajectory of the particles originating from the centre of the circular grid changed to a new point randomly, the standard deviation of the landing points remained stable (i.e. no additional experimental manipulation of expected uncertainty).

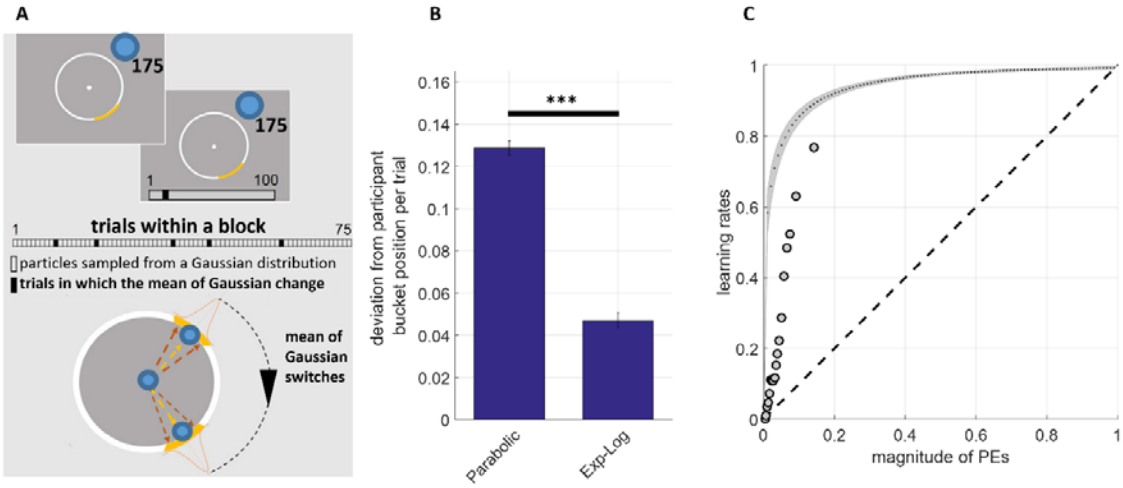


Figure 6. The results from the reanalysis of the bucket task. (A) Participants try to catch blue particles initiating from the centre of a circular grid with a virtual yellow bucket in order to bank as much points as possible. The particle lending trajectories are drawn from a Gaussian distribution with a mean that jumps to a different point randomly (i.e the switch points). Participants need to adjust their inferences in order to place their buckets to where the particles are most likely to land. **(B)** The analysis of the data with the novel reinforcement-learning models demonstrated that the exponential-logarithmic model exhibits greater flexibility to account for participant behaviour (i.e. the model minimising the deviation from participants bucket placement) relative to the cubic model which assumes a strictly parabolic relationship. The values are normalised from 360° to [0,1]. **(C)** In tasks with more precise outcome feedback, the learning rate trajectory is remarkably different relative to traditional reinforcement-learning experiments in which outcomes are binary (i.e. win versus no-win). The model estimates steep learning rate increases over the prediction error space and results mostly align with those reported by Vaghi et al., (2017; average model-free LRs indicated by circular markers).

In this scenario, we show that the exponential-logarithmic model can account for participant behaviour significantly better than the cubic model ($t(48) = 14.168$, $p<.001$). This becomes evident when we plot the trajectory of the estimated LRs over the range of the prediction errors—participants utilise increasingly higher LRs which gradually plateau when the magnitude of prediction errors exceeds 0.2, also revealing a non-parabolic trajectory. These results are widely in agreement

with the initial model-free (i.e. how much participants updated their bucket positions from the previous trial) LRs reported by Vaghi and colleagues (2017, circular data points shown in Figure 6C).

Temporal properties of learning rates

Finally, we conducted our own learning experiment in an independent cohort to investigate the temporal properties of LRs versus prediction errors in a task environment with volatility and noise. Thirty-two healthy volunteers (see Supplementary Table 1 for demographic features) completed a reward magnitude learning task which was adapted and simplified from an affective learning task that served as a control experiment reported in one of our earlier studies[12]. On each trial, participants were asked to choose between two abstract fractals and obtained rewards. The reward associated with the chosen shape was represented as a green filling inside a bar which designated 100 available points per trial. This maximum reward amount was distributed between the options, such that if one shape had more reward points (green filling), the other one would have less (i.e. empty grey area within the bar, Figure 7). Participants completed 4 blocks of 60 trials each, covering a 2 x 2 factorial design of high/low volatility and high/low noise, making the task environment similar to the simulation environment shown in Figure 4.

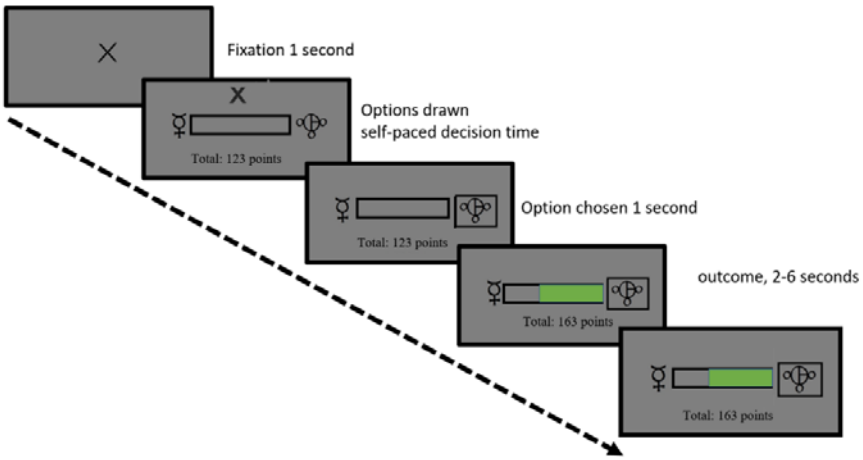
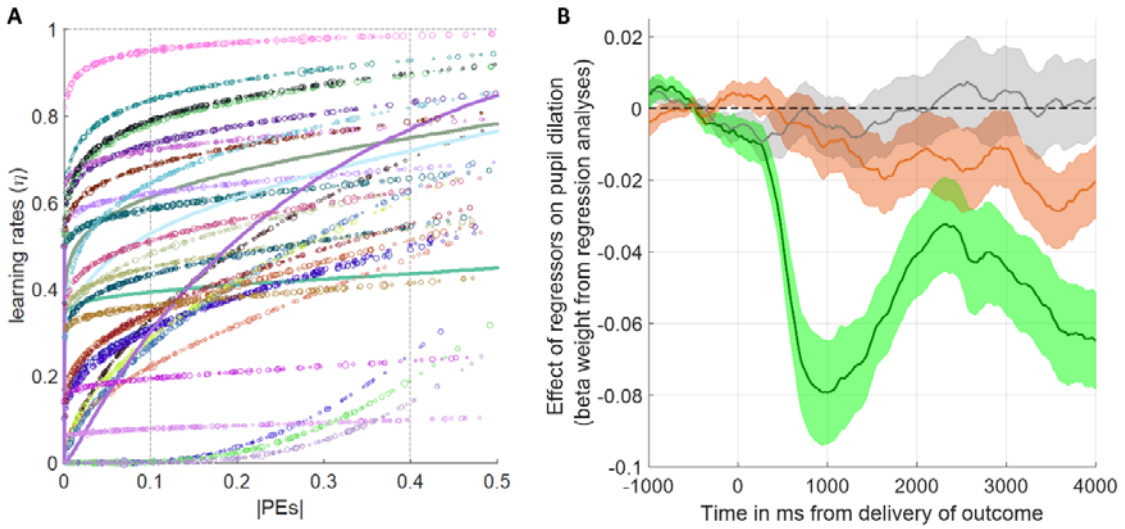


Figure 7. The timeline of the reward magnitude learning task. Following a 1 second fixation period, participants had to choose using the left and right arrow keys between two abstract shapes. They were asked to choose the shape that they thought would give them a larger amount of reward and then instructed to maximise their winnings through trial and error. The full length of the bar represented 100 points. In this trial, the green filling indicates the reward amount for selecting the shape on the right, while the empty (grey) portion of the bar shows the remaining points that could have been won if the unchosen shape was selected. See Supplementary Figure 4 for the task environment, more details about the task available in the Methods section.

Fitting the RW model separately to each task block to assess learning rate adaptation to volatility and noise suggested that participants behaved near-optimally (Supplementary Figure 5A). A formal statistical analysis using a repeated measures ANOVA indicated a significant main effect of volatility ($F(1, 29) = 9.484$, $p=.005$) and a significant interaction between volatility and noise ($F(1, 29) = 5.777$,

364 p=.023). Higher environmental noise reduced LRs in volatile environments (line with information
365 theoretic predictions[20], F(1,29)=7.496, p=.01) but did not influence learning rate estimates in stable
366 environments (Supplementary Figure 5A). As expected, there was a main effect of the type of block
367 participants started the task off with in this noise effect (F(1,29)=4.237, p=.049). Participants' self-
368 reported judgements on environmental volatility and noise, collected at the end of each task block,
369 did not reflect the differences between the conditions, and consequently did not reveal any
370 significant main effects or interaction terms (Supplementary Figure 5B, all p>.280). We reanalysed
371 participant choice behaviour using the exponential-logarithmic model, considering that it
372 demonstrated comparable behaviour with the alternatives as shown in the preceding sections, and it
373 can be fully reduced to the RW model as shown in Figure 3 [and legends] making these models
374 overall easier to compare. The model log likelihoods accounting for how well they fit to participant
375 choice behaviour were not significantly different between the exponential-logarithmic model, RW
376 with 4 LRs (i.e. a separate learning rate for each block) and RW with one learning rate fitted to all
377 blocks all at once (F(2,95)=2.28, p=.108). The relationship between LRs and prediction errors as
378 estimated by the exponential-logarithmic model is shown in Figure 8A and reveals a mixture of
379 trajectories: some resembling a parabolic relationship; some resembling those shown for the data
380 from Vaghi et al., in the preceding section, others similar to the parabolic trajectory shown in the re-
381 analysis of data from Behrens et al., and also RW-like constant with a subtle and gradual linear
382 increase over the PEs. These trajectories further highlight that exponential-logarithmic model can
383 flexibly capture individual variability in RL. The PEs generated under this model demonstrates that in
384 environments with variable [as opposed to binary] outcomes, prediction errors encode the structure
385 of the environment (Supplementary Figure 4).



386

387 **Figure 8. The LRs and their pupillary correlates. (A)** Relationship between the absolute value of prediction errors and LRs.
388 The maximum value on x-axis capped at 0.5, as majority of the participants did not experience prediction errors higher

than this value. Each single-subject is represented as different colours. Variability in individual marker size along each single-subject trajectory is scaled to the average normalised pupil size during the outcome delivery period at that intersection of prediction errors and LRs. Four thick continuous lines designate the participants whose pupil data was corrupted, demonstrating only the behavioural relationship. **(B)** Time evolution of regression coefficients during the outcome delivery period: unsigned prediction errors (grey); LRs (orange); chosen outcome magnitude (green).

Finally, we investigated the temporal properties of the LRs encoded by pupil dilation, along with other key variables such as signed PEs and chosen outcome magnitudes. We used an approach similar to model-based analysis of functional MRI data[31], using an ordinary least squares regression model that we also used in a number of our previous studies[12,32]. Our prediction was that LRs should be instantiated on the same trial, therefore should influence pupil dilation in a gradually increasing manner during the outcome delivery period (i.e. the latter end of the outcome delivery period) while expectations are updated. To capture this effect, we investigated the within-subject correlation between time during the outcome delivery period and regression coefficients for LRs. As shown in Figure 8B, higher LRs (orange line) gradually led to pupil constriction during the outcome delivery period (population average for within-subject correlation coefficient $r(1998)=-.2502$ corresponding to $p<1.57\times10^{-27}$; population average r-values significantly lower than zero, $p=.0249$). Average pupil response encoding LRs was significantly lower than baseline in the final phase of the outcome delivery period ($t(27)=-2.220$, $p=.035$). Here, it is worthwhile to highlight that due to our task design (Figure 7) the chosen outcome magnitudes are represented as bar fillings resulting in greater illumination during the outcome period relative to the fixation cross which marks the beginning of a trial. This is also demonstrated in an even stronger pupil constriction in response to chosen outcome magnitudes (green lines in Figure 8B). Consequently, replication of the pupillary correlates of LRs in different task designs may yield a positive correlation during the outcome delivery period. Nevertheless, a prediction arising from this work would be that the physiological response to LRs should continuously build up during the outcome delivery period as previous expectations are updated in the face of new evidence. Signed PEs encoding the global task structure (Supplementary Figure 4) did not influence pupil dilation (Figure 8B, grey line). We also report results from iterative stochastic generate-recover simulations which indicated desirable parameter recovery metrics (e.g. stable log likelihood values, Supplementary Figure 8A) and demonstrated that the model recovery can recapitulate the behavioural pattern observed in the behavioural experiment (Supplementary Figure 8B).

Discussion

In this work, we showcase two nonlinear update rules (i.e. cubic or exponential-logarithmic) for human reinforcement-learning (RL). One of these models relied on a strong *a priori* parabolic assumption in describing the relationship between LRs and prediction errors (PEs), whereas the other was data-driven and displayed even greater flexibility in capturing the individual variability in human learning behaviour. To the best of our knowledge, these models were not previously reported in the literature. Both simulated data (Figure 4), and analysis of human learning behaviour from different datasets (in Figures 5, 6 and 8) demonstrated that newly proposed models perform comparably well with the existing RL models while generating trial-wise learning rate estimates from participant choice behaviour. The proposed nonlinear models made a concrete and logical prediction about the temporal properties of LRs for the epoch immediately after the decision onset (Supplementary Figure 3), and pupillary data collected during human reward-guided learning seemed to confirm this prediction (Figure 8B).

We think that these nonlinear update rules will have very important implications for neural sciences. Analysis of various functional neuroimaging modalities (e.g. fMRI, MEG, pupillometry, etc.) heavily relies on general linear models (GLMs) which require trial-wise estimates of behavioural processes as predictive variables (i.e. regressors). In GLMs, using decorrelated regressors is vitally important for model efficiency and diminishing collinearity. Consequently, extracting reliable, trial-wise estimates of LRs that are also decorrelated with PEs is an extremely important issue in neural sciences. We showed that both nonlinear models meet this objective. This feature of the novel nonlinear models is also shared by the hybrid Pearce-Hall model (Eqs. 2-3). The LRs generated under these novel models were all highly significantly correlated with those generated under the hybrid Pearce-Hall model, giving the nonlinear models external validity.

To the best of our knowledge no study has independently manipulated the sources of expected and unexpected uncertainties in learning environments (such as the task environment shown in Figure 4 and Supplementary Figure 4), without explicitly informing participants about the properties of the task environment in which they were asked to perform reinforcement-learning. Future studies using our 2x2 (volatility and noise) full factorial approach can probe where different components of RL models are encoded in the human brain more precisely (i.e. striatal regions, habenula [encoding negative PEs][33], bilateral amygdala and dACC), and where this information is integrated in the brain. We think that our more flexible and data-driven exponential-logarithmic model can lead the way, as it can be fully reduced to the RW model (Figure 3) if the relationship between PEs and LRs happen to be constant and linear when estimated from participant choice behaviour. Consequently, we think

that benefits of relying on a 2-parameter exponential-logarithmic model to generate reliable trial-wise learning rate estimates grossly outweigh relying on the traditional 1-parameter RW model. Although we provided evidence to support temporal predictions based on this model (Figure 8B), at this moment in time we were not able to include functional neuroimaging to be able to demonstrate the full neural circuitry associated with it due to funding limitations. However, we think that the neuroimaging directions that we outline here are important not only for cognitive/computational neuroscience but also for promoting a better understanding of psychiatric disorders in which learning processes are impaired.

It might be worthwhile to discuss the merits of the cubic model, separately from the commentary above. This elegantly expressed model (Eq.5) may be particularly useful in tasks in which expected uncertainty (i.e. noise) is experimentally manipulated[34,35] with a finer gradient in magnitude learning tasks, while keeping the average reward amount associated with one shape stable at different levels (e.g. 0.1, 0.2, 0.3, 0.4 …). In our own lab-based experiment (Figure 7 and Supplementary Figure 4), we did not focus on systematically covering the average reward x noise (i.e. SD) space and our choice of 0.5 as average reward rate in stable environments inevitably pushed towards maximum entropy. Experiments systematically covering this stimuli space should be considered for future research. In such experiments, the parameter σ (Eq.2 and Figure 2) can be estimated from participant choice behaviour as the single model parameter which accounts for the participants' inference about the level of noise/expected uncertainty in the task environment. This would help us to understand how participants own understanding of environmental noise as assessed by subsequent self-reports and as parameterised in a learning model map on to actual noise level in the environment. In this preliminary work, we showed that participants did not have a good, higher-order understanding of environmental volatility and noise, despite showing near-optimal learning behaviour (Supplementary Figure 5). The output of our cubic model (i.e. variable, trial-wise LRs) is similar to a recently described neurophysiological reward-dependent metaplasicity model that can account for the learning behaviour of non-human primates[36]. However, under the cubic model the level of expected uncertainty in the environment determines the *transition gradient* with which agents should attribute higher PEs to unexpected uncertainty/volatility (Figure 2), instead of expected uncertainty setting a binary *detection baseline* for changes which should be attributed to environmental change[19]. Needless to highlight, the cubic model is much easier implement on data collected from human volunteers due to feasibility limitations in electrophysiological recording. Overall, although our recommendation for learning scholars would be to rely on the exponential-logarithmic model, the cubic learning rule can also be a viable alternative as it is simple, intuitive and can extract trial-wise LRs and PEs which are decorrelated with each other.

In the past two and a half decades, cognitive computational neuroscience have made big advances by studying the neural correlates of PEs. Consequently, the number of studies that investigated brain regions encoding PEs significantly outnumber those that investigated neural regions encoding LRs. One study previously described in this paper using the hybrid Pearce-Hall model described in this paper in a reversal-learning task involving aversive outcomes (a stable 30% probability of receiving an electroshock) showed that LRs are encoded in the human amygdala, bilaterally[14]. On the other hand, two key studies that manipulated the volatility (i.e. the unexpected uncertainty) of learning environments showed that volatility of the environment, which in return correlates significantly with LRs, are encoded in the dorsal anterior cingulate cortex (dACC) in both non-social and social learning environments[4,16]. Recently, we have reviewed learning and value-based decision-making literature across rodent, monkey and human imaging studies and our synthesis also pointed out to a role for dACC potentially guiding LRs[33]. We hope that models that we communicate in this manuscript will ignite further interest in exploring the neural correlates of LRs as opposed to PEs, and in doing so, we will help to define the next decades of functional neuroimaging and electrophysiology research.

Methods

Reanalysis of existing datasets. In the reanalyses of behavioural data from Behrens et al., (2007) a maximum likelihood approach was used, which was implemented by running MATLAB's *fmincon* function iteratively with random starts covering the whole parameter space (1000+ iterations per participant per condition). No trials or participants were excluded from the analysis. In the reanalysis of behavioral data from Vaghi et al., (2017) parameters were estimated in a similar manner to maximum likelihood approach (i.e. identifying the parameter combination that minimises the delta between model estimates and participants' actual bucket positions in the normalised space).

All model and analysis scripts are available from an Open Science Framework repository: https://osf.io/g7mkv/?view_only=5b4c2ec0eb8442bf95fd26e151cd71d0

Pupillometry Experiment. The study involved a single experimental session during which participants completed a novel reward learning task. The task was displayed on a VGA monitor. Participants' heads were stabilized during the task while an eye-tracking system (Eyelink 1000 Plus; SR Research, Ottawa, Canada) was recording pupillometry. The pupil dilation was recorded at a sampling rate of 500 Hz. The VGA monitor was linked to the laptop running the experiment with Psychtoolbox on MATLAB.

The task consisted of 4 blocks of 60 trials each. The decision time was self-paced, with additional rest period after every block. Participants were presented with two abstract shapes from the Agathodaimon font (Figure 7). The abstract shapes of the learning task were offset by around 7˚ visual angle and were displayed on either side of an empty black score bar as well as a fixation cross indicating the middle of the screen during the inter-trial interval. The shapes were presented randomly on each side of the screen, prompting participants to identify the better shape rather than right versus left side of the screen. The outcome reward magnitude following participant choice was displayed inside the empty bar between the options, as a proportional green filling for a jittered interval of 2–6 seconds. The shapes were the same for each trial during a given block but changed into the new block (total of 2x4 shapes). At the end of each trial, participants' running total was updated on the screen. The order in which all the blocks were completed was counterbalanced across participants. The information content of the reward outcomes was varied across the four blocks by altering the volatility and noise associated with the stimulus-outcome pair. The outcomes for each block were thus generated following a 2 x 2 factorial of high/low volatility and high/low noise schedule (Supplementary Figure 4). For the entire task, the quantity and variety of stimuli remained constant in order to minimise disparities in luminance between trials. At the end of the experiment, the outcome of 20 randomly selected trials were converted to real money which was

553 added to a baseline payment of £10. The participants were told to bank as many points as possible
554 at the beginning of the experiment. After each block, participants were asked to rate the number of
555 times they thought the average reward shifted from low to high, and how wide they thought the
556 range of rewards were around the average reward value (Supplementary Figure 5B).

557 **Pupillometry Preprocessing and Analysis.** We followed identical steps to our previous eye tracking
558 studies in preprocessing and analysing pupillary data[12,32]. Pupillometry data was collected to
559 examine the participants' physiological responses during the reward-guided learning task. Eyeblinks
560 were removed from the data by using the Eyelink system's built-in filter. This was followed by a
561 linear interpolation to compensate for missing data points. The resulting pupillary trace was
562 processed through a low pass Butterworth filter with a cut-off of 3.75 Hz, and then z-transformed
563 across the experimental session[37]. The pupillary correlates of the model-based regressors were
564 assessed by fitting a multiple linear regression model on baseline corrected pupil data. If over half of
565 the data from the outcome period had been interpolated, the affected trial was excluded from the
566 analysis. Participants who had missing data more than 50% of the trials were excluded from the
567 analysis, leading to the exclusion of four cases in total.

568 The pupillary multiple linear regression model had the following regressors: an intercept, trial
569 number (as a proxy for tiredness), chosen reward outcome, PEs and LRs.

570 **Acknowledgements**

571 The authors would like to thank to Prof. Tim Behrens and Dr Matilde Vaghi for sharing the
572 behavioural data and a helpful discussion. The authors would also like to thank to Prof. Peter Dayan,
573 Prof Michael Browning and the participants of the Oxford Computational Psychiatry Journal Club for
574 their helpful comments about behavioural modelling. EP is full-time employed on a grant jointly
575 funded by the UK Medical Research Council and Janssen Pharmaceuticals (MR/S035591/1) awarded
576 to Catherine J Harmer, Susannah Murphy and Philip J Cowen. BI is supported by Rhodes Scholarship
577 for West Africa. The pupillometry study was funded by the University of Oxford Department of
578 Psychiatry awarded to BI and EP.

579

580

581

582

583

584 **References**

585 1 Angela, J. Y. & Dayan, P. Uncertainty, neuromodulation, and attention. *Neuron* **46**, 681-692
586 (2005).
587 2 Erev, I. & Roth, A. E. Predicting how people play games: Reinforcement learning in
588 experimental games with unique, mixed strategy equilibria. *American economic review*, 848-
589 881 (1998).
590 3 Sutton, R. S. & Barto, A. G. *Introduction to reinforcement learning*. Vol. 135 (MIT press
591 Cambridge, 1998).
592 4 Behrens, T. E., Hunt, L. T., Woolrich, M. W. & Rushworth, M. F. Associative learning of social
593 value. *Nature* **456**, 245-249 (2008).
594 5 Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward.
595 *Science* **275**, 1593-1599 (1997).
596 6 O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H. & Dolan, R. J. Temporal difference models
597 and reward-related learning in the human brain. *Neuron* **38**, 329-337 (2003).
598 7 Franklin, N. T. & Frank, M. J. A cholinergic feedback circuit to regulate striatal population
599 uncertainty and optimize reinforcement learning. *Elife* **4**, e12029 (2015).
600 8 Khorsand, P. & Soltani, A. Optimal structure of metaplasticity for adaptive learning. *PLoS*
601 *computational biology* **13**, e1005630 (2017).
602 9 Silver, D. *et al.* Mastering the game of go without human knowledge. *Nature* **550**, 354
603 (2017).
604 10 Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *nature*
605 **529**, 484 (2016).
606 11 Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X. & Bishop, S. J. Anxious individuals
607 have difficulty learning the causal statistics of aversive environments. *Nature neuroscience*
608 (2015).
609 12 Pulcu, E. & Browning, M. Affective bias as a rational response to the statistics of rewards and
610 punishments. *Elife* **6** (2017).
611 13 Rescorla, R. A. & Wagner, A. R. A theory of Pavlovian conditioning: Variations in the
612 effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current*
613 *research and theory* **2**, 64-99 (1972).
614 14 Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A. & Daw, N. D. Differential roles of human
615 striatum and amygdala in associative learning. *Nature neuroscience* **14**, 1250 (2011).
616 15 Pearce, J. M. & Hall, G. A model for Pavlovian learning: variations in the effectiveness of
617 conditioned but not of unconditioned stimuli. *Psychological review* **87**, 532 (1980).
618 16 Behrens, T. E., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. Learning the value of
619 information in an uncertain world. *Nature neuroscience* **10**, 1214-1221 (2007).
620 17 Mathys, C., Daunizeau, J., Friston, K. J. & Stephan, K. E. A Bayesian foundation for individual
621 learning under uncertainty. *Frontiers in human neuroscience* **5**, 39 (2011).
622 18 Roesch, M. R., Esber, G. R., Li, J., Daw, N. D. & Schoenbaum, G. Surprise! Neural correlates of
623 Pearce–Hall and Rescorla–Wagner coexist within the brain. *European Journal of*
624 *Neuroscience* **35**, 1190-1200 (2012).
625 19 Soltani, A. & Izquierdo, A. Adaptive learning under expected and unexpected uncertainty.
626 *Nature Reviews Neuroscience*, 1 (2019).
627 20 Pulcu, E. & Browning, M. The misestimation of uncertainty in affective disorders. *Trends in*
628 *cognitive sciences* (2019).
629 21 Pulcu, E. *et al.* Temporal discounting in major depressive disorder. *Psychological Medicine*, 1-
630 10 (2013).
631 22 Murphy, J., Vuchinich, R. & Simpson, C. Delayed reward and cost discounting. *The*
632 *Psychological Record* **51**, 571-588 (2001).

23 Kahneman, D. & Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263-291 (1979).
24 Stott, H. P. Cumulative prospect theory's functional menagerie. *Journal of Risk and uncertainty* **32**, 101-130 (2006).
25 Daw, N. D. Trial-by-trial data analysis using computational models. *Decision making, affect, and learning: Attention and performance XXIII* **23**, 1 (2011).
26 Luce, R. D. The choice axiom after twenty years. *Journal of mathematical psychology* **15**, 215-233 (1977).
27 Wang, J. X. *et al.* Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience* **21**, 860 (2018).
28 Nassar, M. R., Wilson, R. C., Heasly, B. & Gold, J. I. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience* **30**, 12366-12378 (2010).
29 Prelec, D. The probability weighting function. *Econometrica*, 497-527 (1998).
30 Vaghi, M. M. *et al.* Compulsivity reveals a novel dissociation between action and confidence. *Neuron* **96**, 348-354. e344 (2017).
31 Behrens, T. E., Hunt, L. T. & Rushworth, M. F. The computation of social behavior. *science* **324**, 1160-1164 (2009).
32 Murphy, D. A., Xie, J., Harmer, C. J., Browning, M. & Pulcu, E. Dynamic modulation of inequality aversion in human interpersonal negotiations. *Communications biology* **5**, 1-14 (2022).
33 Pulcu, E., Guinea, C., Cowen, P. J., Murphy, S. E. & Harmer, C. J. A translational perspective on the anti-anhedonic effect of ketamine and its neural underpinnings. *Molecular Psychiatry*, 1-7 (2021).
34 Diederen, K. M., Spencer, T., Vestergaard, M. D., Fletcher, P. C. & Schultz, W. Adaptive prediction error coding in the human midbrain and striatum facilitates behavioral adaptation and learning efficiency. *Neuron* **90**, 1127-1138 (2016).
35 Diederen, K. M. & Schultz, W. Scaling prediction errors to reward variability benefits error-driven learning in humans. *Journal of Neurophysiology* **114**, 1628-1640 (2015).
36 Farashahi, S. *et al.* Metaplasticity as a neural substrate for adaptive learning and choice under uncertainty. *Neuron* **94**, 401-414. e406 (2017).
37 Nassar, M. R. *et al.* Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature neuroscience* **15**, 1040-1046 (2012).
38 Pulcu, E. *et al.* Using a generative model of affect to characterize affective variability and its response to treatment in bipolar disorder. *Proceedings of the National Academy of Sciences* **119**, e2202983119 (2022).
39 Nassar, M. R. *et al.* Age differences in learning emerge from an insufficient representation of uncertainty in older adults. *Nature Communications* **7**, 11609 (2016).

677 ## Supplementary Tables and Figures

678 **Table 1: Demographic details of participants**

Measure	Mean (SD)
Age	29.4 (9.79)
Gender	56.25% Female
Years of Education	17.72 (2.00)
State-STAI	30.13 (7.36)
Trait-STAI	35.88 (7.61)
QIDS-16	6.22 (4.33)

679

680 QIDS-16; Quick Inventory of Depressive Symptoms, 16 item self-report version. Trait/State-STAI; Spielberger State Trait
681 Anxiety Inventory. Note that scores of 6 or above on the QIDS-16 indicate the presence of mild depressive symptoms.
682 The Trait/State-STAI has no standard cut-off scores.

683

684

685

686

687

688

689

690

691

692

693

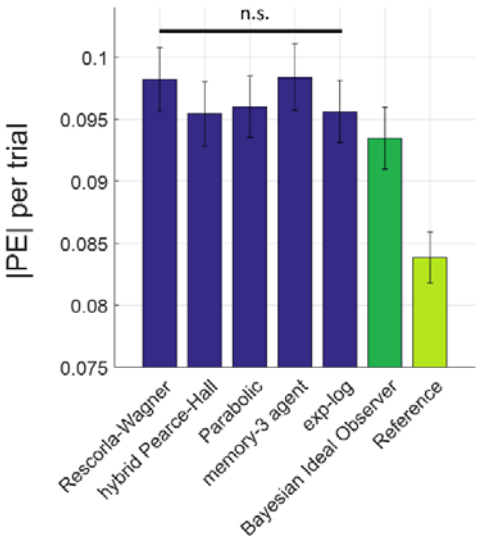
694

695

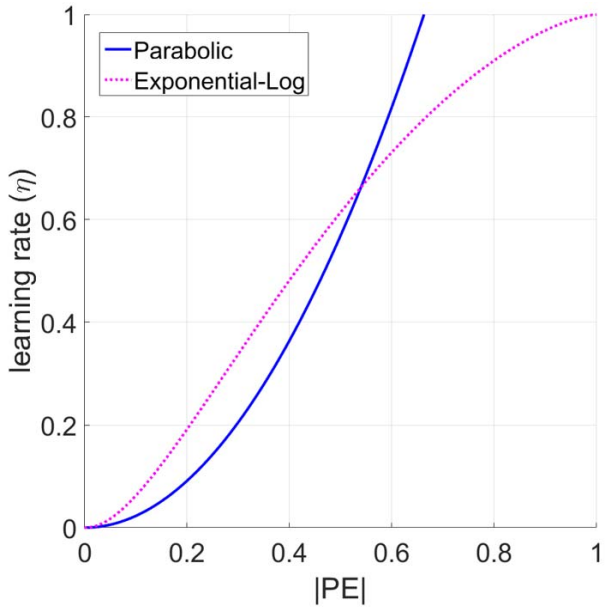
696

697

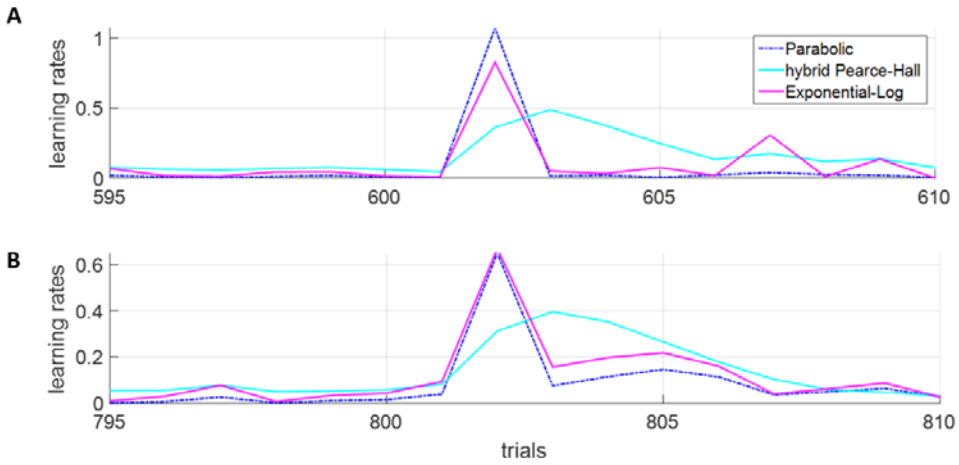
698



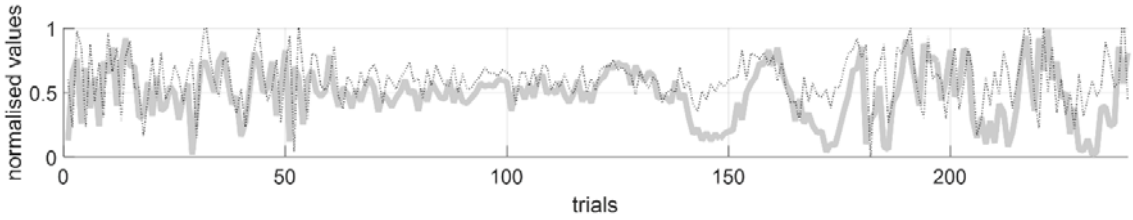
Supplementary Figure 1. The average magnitude of prediction errors (|PE| on y-axis) under different learning models. Across 1500 trials in the simulated task environment (Figure 4), the RL models lead to average PEs which were not statistically significantly different. Based on an exponential relationship between environmental noise and the free-parameter κ (Figure 2), any memory-n agent (i.e. computing the expected uncertainty from the standard deviation of outcomes observed in the environment based on a history of n previous trials) can adjust its learning behaviour dynamically. An example is shown for a memory-3 agent who computes the standard deviation of outcomes based on the observations in the past 3 trials. The behaviour of a recently described 5-dimensional Bayesian Ideal Observer model[20,38] is shown for illustrative purposes (dark green bar). The "Reference" bar (light green) refers to |PE|s computed with respect to the mean of the generative process (i.e. if an agent had all the information about the environment up front).



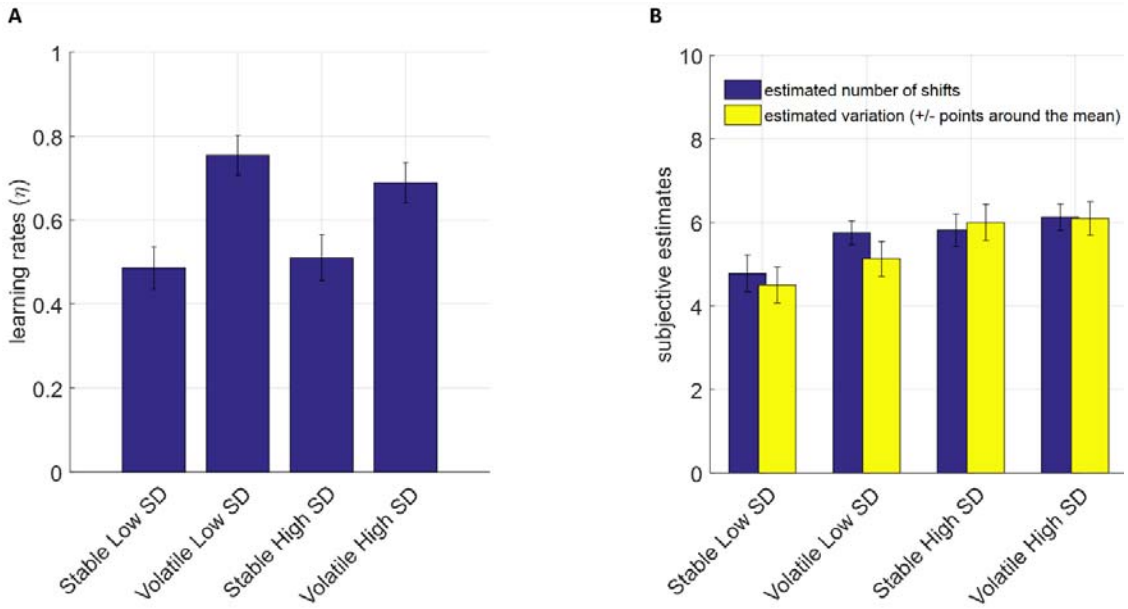
Supplementary Figure 2. A nonlinear relationship between the absolute magnitude of prediction errors (|PE|) and LRs. The exponential-logarithmic model which behaves marginally, but not statistically, better in the simulation environment (Figure 4) have steeper LRs relative to the cubic model for observations leading to PEs ≤ 0.50. The maximum PE magnitude (i.e. the largest difference between any n-(n-1) outcomes) was 0.71 in the simulation environment. Therefore, the trajectory of the curves above |PE|>0.71 on x-axis is theoretical.



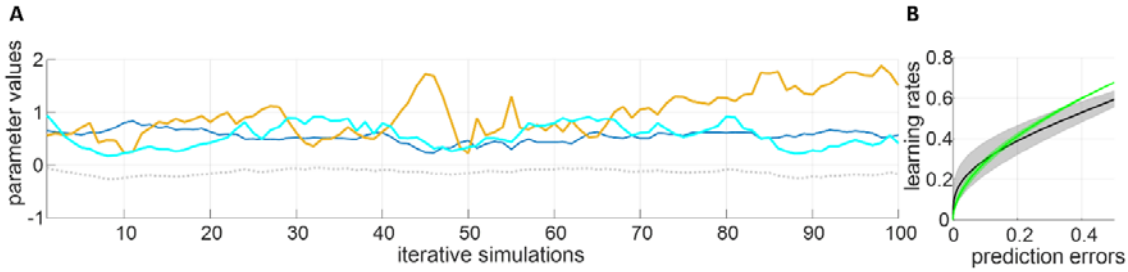
Supplementary Figure 3. Learning rate acceleration between hybrid Pearce-Hall, cubic and exponential-logarithmic models. The hybrid Pearce-Hall model is slower to accelerate and decelerate LRs. The trial numbers on x-axes refer to the task environment shown in Figure 4. Examples are shown for **(A)** transitions within high volatility high noise, and **(B)** between high volatility high noise and low volatility low noise periods. In tasks designs with explicit presentation of outcomes (e.g. Figure 4 in which exact outcome magnitudes are shown in the task environment) LRs can marginally be higher than 1[37,39].



Supplementary Figure 4. Task environment of the reward magnitude learning task (thick grey line) as experienced by a participant. Block orders were randomised. Prediction errors from the exponential-logarithmic model superimposed (thin dashed black line), demonstrating that prediction errors globally encode the structure of the environment.



Supplementary Figure 5. Behavioural results. (A) LRs measured by the RW model fitted to each block separately, demonstrate an overall normative behaviour in human learners. **(B)** Participants' self-evaluation of environmental volatility and noise did not reflect the changes in the task structure. Estimated number of shifts aiming to capture volatility in the environment, whereas variation around the mean aimed at capturing participants' inference about environmental noise.



Supplementary Figure 6. Generate-Recover simulation results. (A) Initiating from population mean parameter values of the exponential-logarithmic model (delta: cyan; gamma: navy; beta: orange line) in a randomly selected task environment, 100 iterative stochastic generate recover procedures were conducted (i.e. choices were generated stochastically based on initial parameter values and parameters were re-estimated based on newly generated choices). If a deterministic procedure is used (e.g., chose left if choice probability for the left shape is higher than 0.5) the parameter estimates would flat-line from the third simulation onward. Log likelihood of the model (dashed grey line at the bottom) remains at a very desirable range throughout 100 simulations. Beta is linearly transformed by dividing the values by 100 for plotting purposes. **(B)** The relationship between learning rates and prediction errors plotted (green line and error shading) based on the mean parameter values from the generate-recover simulations shown in panel A. Black line and grey error shading denotes the population mean of the behavioural experimental results reported in Figure 8A.