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, and the predictive significance of grey clouds may change from season to season). Living in an uncertain world continuously recruits learning systems in the brain, with which agents try to make better decisions in their environments. Behavioural adaptation to dynamically changing environments is a type of value-based decision-making which 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[2-4] and neural sciences[5-10], as well as machine learning[11,12].

In cognitive neuroscience of RL, a wide majority of the literature focuses on neural/physiological correlates of prediction errors, whereas the relationship between prediction errors and learning rates, and generally the neural/physiological correlates of learning rates remains mostly understudied[13]. We think this is a byproduct of a lack of approachable models which would allow investigation of the relationship between prediction errors and learning rates, instead of a lack of interest in this relationship/research question. In this paper, we describe two simple RL algorithms that can adequately address this knowledge gap. We describe the properties[14] of these simple update rules, and test their applicability to different experimental tasks.

Let us start by considering the factors which influence learning rates, as evidenced by literature that utilised existing RL algorithms. A number of previous studies showed that human learners adjust their learning rates (LRs) based on the information content of outcomes[15,16], utilising a higher learning rate for outcomes more informative for predicting the future state of the environment. In computational neuroscience, LRs are commonly modelled by using different RL algorithms[13,17,18] — which are in effect approximations of the Bayesian models[19,20] that can also estimate otherwise hidden generative rules of dynamically changing environments. The Rescorla-Wagner (RW) algorithm is undisputedly the most widely used RL model[17] in the literature:

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

where $\tilde{p}$ is the agent's probability estimate for an outcome, letter ***O*** designates 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, to be able to estimate participants' average LRs during that period. This analytic approach uses the RW model as an assessment tool to measure the speed of learning, implemented by a number of key RL

studies[5,15,16,19] 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, or do not change very frequently. Similarly, LRs also change [nonlinearly] as a function of whether PEs are positive or negative[21,22]. Globally, these findings clearly indicate that LRs are not a constant coefficient. 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[13,18,23] 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):

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

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 ($|\,(O_{(t)} - \tilde{p}_{(t)})\,|$) is commonly used as a proxy for the degree of surprise associated with observing the outcome in question. The model would then update the agent's estimates of the reward probability into the next trial:

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

Where λ is another free-parameter commonly used for smoothing the effect of the trialwise learning rates which are updated by the surprise signal, therefore finetuning the agent's estimation of reward probability. 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 the speed of learning which would be optimal in a given environment[1,24,25]. *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. In the literature expected uncertainty is also referred to as aleatoric uncertainty[26], environmental noise[25] or stochasticity[27]. 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[1,28]. Surprise signals also decay in environments with high expected uncertainty, as prediction errors arising from violation of one's expectations become less informative. 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 low learning rate that reflects the limited epistemic utility of observations for fine-tuning predictions. However, there can be instances when learned associations are rendered out-of-date relative to the underlying relationships that currently prevail in the environment, since these relationships can change systematically over time, or change suddenly for a shorter period of 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*"[19]). 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. A real-life example could be a sudden increase in UK daily temperatures way above the seasonal average, brought along by an Atlantic storm. 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 state of the environment in the near future. 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.

Historically, learning under expected and unexpected uncertainty has also been investigated by Bayesian models[19,20,27], including our own[25,29], and their approximations[30,31]. These generative models are particularly useful for delineating the true source of environmental uncertainty (i.e. whether it is attributable more to expected or unexpected changes). Although these models are important to mention, they are often complex (i.e. despite some of them being openly available online, they can be regarded as "black box" algorithms for the wide majority of users) and not necessarily plausible representations of computations underlying human learning. For example, emerging evidence from our lab suggests that, at best, human learners behave similarly to a

degraded/lesioned Bayesian learner[32]. Furthermore, to the best of our knowledge, there is no consensus in the literature as to how these models should be fitted to human choice behaviour. For example, in their seminal work Behrens and colleagues [from whom we obtained original data and re-analysed below] used a Bayesian Ideal Observer model to create regressors that track environmental volatility for neuroimaging analysis, but did not fit this model to participant choice behaviour as the latter would require making many [arbitrary] assumptions about whether these models should have similar risk or probability weighting preferences as human participants or whether they should also act optimally in these behavioural domains, or how to perform model comparison considering these models do not necessarily have free parameter in the traditional sense, etc. Consequently, it is clearly beyond the scope of the current work to reconcile the differences between these complex models and to come up with a single unifying model that can explain human choice behaviour the best. Here, instead, we focus on simple and relatable learning algorithms that can perform dynamic learning rate update based on a nonlinear relationship between LRs and the magnitude of the PEs. We think the models that we explore in the subsequent sections can empower cognitive scientists, psychologists and clinicians with simple, relatively easier-to-implement tools to allow investigation of questions surrounding learning rates during RL.

A parabolic relationship between learning rates and prediction errors

In RL tasks with probabilistic contingencies and binary outcomes (i.e. reward versus no reward), the PEs are always between -1 and 1. For example, if one option available in the environment returns no rewards (i.e. a vector of null outcomes: [0, 0, 0, 0 ...]) yet the agent continues to bet on that option thinking their luck will turn, the lowest values possible for the PEs experienced during that period can be is a vector of -1s (i.e. if the agent remains implacably confident throughout that the outcome surely must occur on this next occasion). 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 outcome events 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[33,34], risk perception[35,36], stochastic choice[37,38] etc.) across human and non-human primates, a simple functional form satisfying all of these assumptions proposes a parabolic relationship between the PEs and LRs:

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

In this notation 4 in the denominator ensures that the focal point of the parabolic curve on the *xy* plane is always (0, κ). The focal point of the parabola adjusts how steeply PEs transform to LRs (Figure 1), and it is the free parameter to be estimated from participant choices.

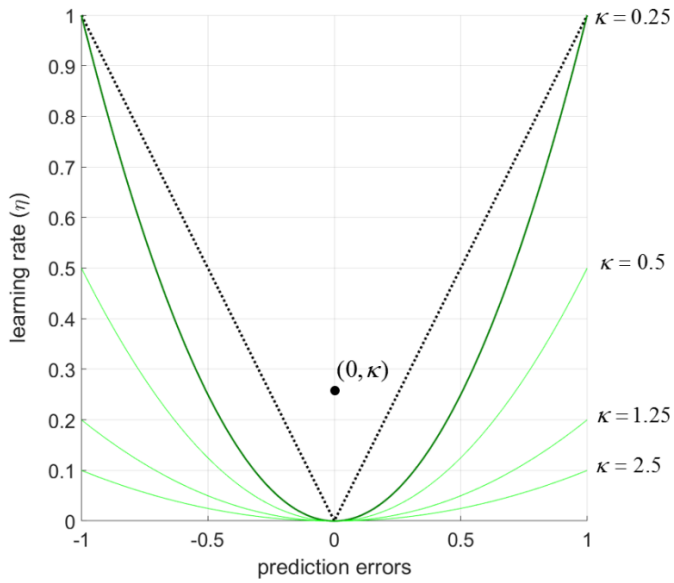


Figure 1. A parabolic relationship between prediction errors and learning rates obtained by a cubic learning algorithm. In environments with high unexpected but low expected uncertainty, one would assume a somewhat linear relationship between prediction errors and LRs (in a V shape designated by the black dashed lines), meaning that agents should update their beliefs proportionately to how far away their beliefs about the environment were 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 for 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, by combining Eqs. 1 and 4, the update algorithm (i.e. updating PEs with corresponding LRs on each trial) can be expressed as:

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

We will refer to this model as the cubic model in the rest of the manuscript. Here, the free-parameter κ cognitively 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 they encounter any unexpected environmental changes (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. 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) characterising the environment[24]. A recent analysis of normative learning models demonstrated that LRs decay exponentially in noisy environments (i.e. those with high expected uncertainty)[39], as expected uncertainty erodes the information content of outcomes. 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:

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

which would allow the model to tune LRs with respect to the magnitude of prediction errors at a given level of environmental noise. Note that in this formulation *a* designates the initial level of expected uncertainty in the environment ($a \neq 0$). Under these assumptions, the nonlinearity of LRs guiding behavioural adaptations could resemble Figure 2, below. The distribution of these simulated LRs is globally in line with empirical 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[30].

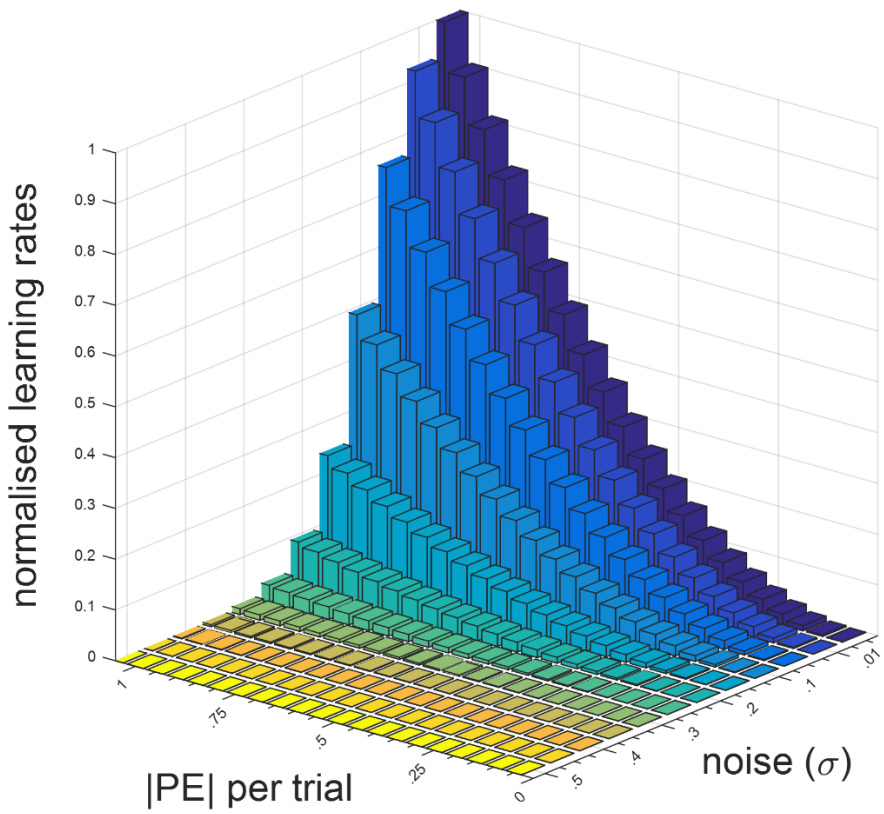


Figure 2. Nonlinearity in trial-wise LRs as a function of prediction errors and environmental noise. Increasing levels of estimated expected uncertainty (σ) diminish the relationship between PEs and LRs. In these situations, the agents should no longer attribute high magnitude prediction errors as indicating informative changes in the environment, but should instead simply treat them as noisy fluctuations which are not informative. These ideas 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 $\kappa = f(\sigma) = \frac{\pi e^{7\pi\sigma}}{10^3} + \frac{\gamma}{6}$. This would suggest that the parameter K is a function of environmental noise, and one example of this exponential relationship is shown in Supplementary Figure 1. In learning tasks in which the expected uncertainty is experimentally manipulated, it may be possible 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. ϒ 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[14]. 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). These parameters work in tandem to reveal the trajectory of the weighting curve. 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). This model describes that agents will tune their LRs in response to different PE magnitudes. The flexibility of this non-linear model allows LRs to slow down or ramp up in response to environmental demands as it can take many different trajectories. It is worthwhile to highlight that cognitively this model serves a similar function in tuning LRs to the magnitude of the PEs that we described above for the cubic model. Therefore, it is possible to conceptualise this model as a more flexible variant of the cubic model without any *a priory* [parabolic] assumptions. Furthermore, the model can also fully reduce to the traditional RW model, a quantitative feature which is not available in the cubic model, as the model is able to generate constant LRs over the prediction errors, as shown by horizontal lines in Figure 3. In our own evaluation, these features make the exponential-logarithmic model better than the cubic model.

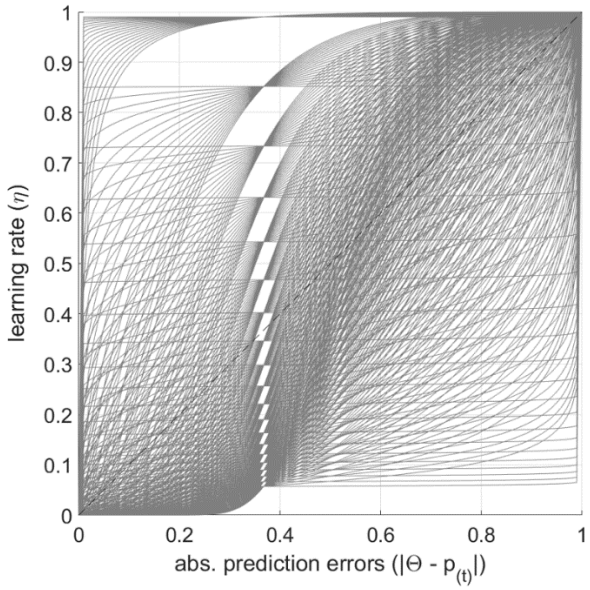


Figure 3. Nonlinearity in trial-wise LRs as a function of prediction errors under an exponential-logarithmic assumption. An exponential-logarithmic weighting function allows greater flexibility in capturing individual variability in how PEs shape LRs. The parameters δ and λ are coarsely sampled for illustration 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.

In the subsequent sections, we simulate the behaviour of these cubic and exponential-logarithmic models to evidence how they can flexibly adapt to reward outcomes in dynamically changing environments, followed by demonstrating their performance in RL tasks with binary and continuous/gradient outcomes.

Results

Simulated data. First, we aimed at testing the upper boundary of how well simulated agents could perform in dynamically changing environments under different learning models. This could be understood as a metric of model flexibility, which is an important aspect of model identification that can describe adaptive learning behaviours. We wanted to test whether the novel models that we proposed can behave on-par with traditional models like the RW model. 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 simulated task environment (from all 1500 trials all at once). 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. Our approach is akin to a maximum likelihood estimation from a search of viable discrete parameter values, as minimising the PEs under a given RL model indicates how well the model-agents' beliefs about the reward rate can converge to the mean of the actual outcome distribution. In these simulations the effect of choice stochasticity was not considered as choices were not simulated with arbitrarily selected learning rate values (which would in effect be a theoretical equivalent of a generate-recover simulation).

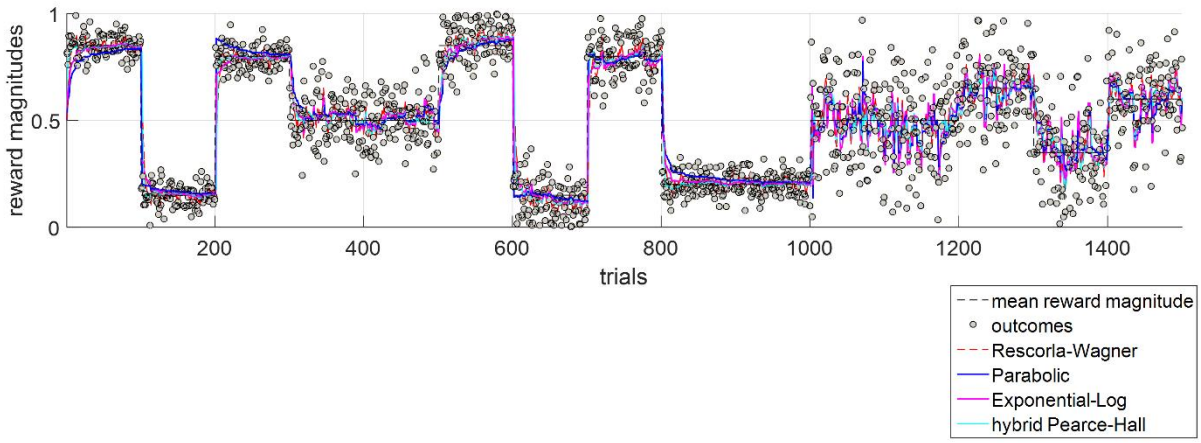


Figure 4. The behaviour of different reinforcement-learning models in a task environment in which unexpected and expected uncertainties were independently manipulated. On each trial, the task environment returns a reward outcome between 0 and 1 (y-axis). The outcomes were generated to be variable from 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). Because models perform ever so comparably, their differences are illustrated in Supplementary Figure 2, showing the average prediction error values relative to the simulated outcomes in the task environment.

In the simulated task environment, all models converge reasonably well with the true average reward rate (different coloured lines designating different models globally overlap in Figure 4). In line with this qualitative/visual evaluation, the magnitudes of prediction errors under each model were not statistically different ($F(4, 7495) = 0.32$, $p = 0.863$, Supplementary Figure 2), meaning that all models can perform behavioural adjustments comparably well. The nonlinear relationship between PEs and LRs under the cubic, the exponential-logarithmic and the hybrid Pearce-Hall models are illustrated in Supplementary Figure 3.

The hybrid Pearce-Hall, the cubic and the exponential-logarithmic models can estimate LRs on trial-by-trial basis. The PEs and LRs estimated under each model (based on model behaviours shown in Figure 4) were decorrelated ($-.01 < r(1498) < .01$, all $p > .80$). However, LRs estimated under the hybrid Pearce-Hall model (mean±SD = 0.096 ± 0.071) were significantly higher than the ones estimated

under the cubic (mean±SD = 0.042 ± 0.094) and exponential-logarithmic models (mean±SD = 0.076 ± 0.12, F (2, 4497) = 119.07, p<.001). Nevertheless, these learning rate estimates from different models were very significantly correlated (all r (1498)>.68, p<.0001). Taken together, these results indicate that 1-parameter cubic and 2-parameter exponential-logarithmic models can be viable alternatives to the existing RL models. Here, it is worthwhile to highlight that one of the main differences between these models is how quickly they would adjust their behaviours in response to environmental change. Under the cubic and exponential-logarithmic models, LRs accelerate as soon as there is a dramatic environmental change, therefore leading to large PEs, and LRs subsequently decay as the magnitude of PEs get smaller. Comparatively, under the hybrid Pearce-Hall model acceleration and deceleration of LRs happen relatively more gradually. This is because LRs here are weighted integration of LRs and the PEs from the previous trials (Eq. 3, Supplementary Figure 4), as determined by a weighting constant (Eq.2) that is estimated from participant choices. Significantly, although each of these three models (Supplementary Figure 3) describe an efficient way of interacting with the environment (i.e. accelerating LRs only in cases in which the changes in the environment are informative), notably, they make different predictions about the temporal properties of LRs (Supplementary Figure 4). In the hybrid Pearce-Hall model the information is fed forward as the learning rate is a secondary process, one which is updated in parallel (Eqs. 2-3) in addition to estimated reward rate. In contrast, the cubic and exponential-logarithmic models both assume that incoming prediction errors would trigger their corresponding LRs instantaneously during the feedback period of the current trial. We tested these divergent predictions on temporal properties of LRs subsequently in a novel RL experiment with pupillometry.

Human behaviour. First, we investigated how well nonlinear RL models explain human behaviour in two previously published data sets. The first dataset contained healthy volunteers (N=18) who completed a RL task with stable (120 trials) or volatile (170 trials) reward contingencies in different task blocks[19]. In the original work by Behrens and colleagues[19], participants were asked to choose between two fractals which returned variable reward amounts on each trial or a null outcome, based on the underlying reward probabilities that participants needed to learn. In the learning task, the reward magnitudes associated with each fractal were generated randomly. In the current reanalysis of this data, the stochastic choice/sigmoid function fitted to participant choice behaviour was identical across all RL models. More specifically, participants' decision values were computed by integrating their reward probability estimates (computed under each RL model) with their corresponding reward magnitudes associated with each option. The reward magnitudes (m) associated with each shape were modulated by a single power utility parameter (ρ) accounting for participants' risk attitude[35], such that $\tilde{m} = m^{\rho}$. Then, the expected value difference between two

options was calculated and fed into a sigmoid function to generate the choice probabilities[37]. This is a very common approach in modelling participant choice behaviour in this task.

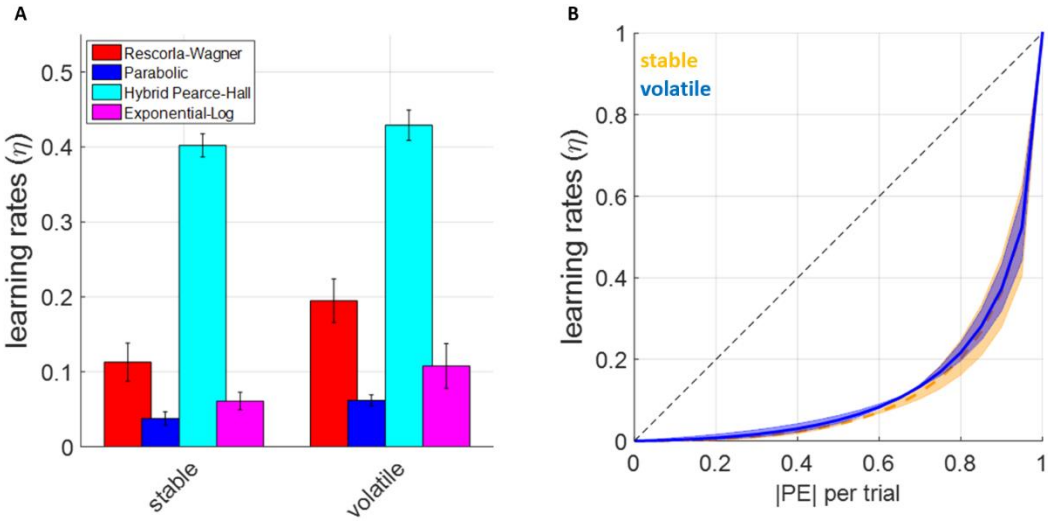


Figure 5. The LRs estimated under each reinforcement-learning model. (A) Average LRs estimated under each model. Hybrid Pearce-Hall model estimates higher LRs relative to the Rescorla-Wagner model, whereas cubic and exponential-logarithmic models estimate overall lower LRs. **(B)** The relationship between PEs and LRs estimated by the exponential-logarithmic model reveals a nonlinear trajectory very similar to what is proposed by the cubic model. Volatility of different task blocks did not seem to influence learning rate trajectories in terms of how absolute values of the PEs tended to influence LRs. This relationship was similarly exponential in form across both stable and volatile task-blocks, as shown.

Overall, all RL models estimated higher LRs in the volatile relative to the stable block, replicating the findings previously reported by Behrens and colleagues (2007). Analysis of model log likelihoods by fitting a 2x4 repeated measures analysis of variance (ANOVA) did not reveal any significant main effect of block type (i.e. volatile or stable; $F(1,68)=.014$, $p=.907$) or model ($F(1,68)=.018$, $p=.997$), or any significant block type x model interaction ($F(3,68)=.002$, $p>.999$). Considering that the RW and cubic models (one), and hybrid Pearce Hall and exponential logarithmic models (two) had identical number of learning parameters we performed subsequent model comparisons within each model group. This analysis suggested that the models were all statistically comparable (log likelihood values are only differentiated at the 3rd decimal place, $p>.923$) and all models tested here fit both volatile and stable environments comparably well. Bayes Omnibus risk calculated for these comparisons is 0.943 indicating that it is difficult to discriminate between model fits.

The difference in LRs between stable and volatile blocks were only statistically significant under the RW and cubic models ($t(34)$ values 2.114 and 2.079, p values 0.042 and 0.045, respectively). The average LRs from the hybrid Pearce-Hall and the exponential-logarithmic models did not capture a statistically significant volatility effect ($t(34)=1.057$ and 1.460, $p=.298$ and .153, respectively). Similar to the simulated results reported in the preceding section, the hybrid Pearce-Hall model generated much higher learning rate estimates relative to other models (Figure 5A). The analysis

with the cubic model indicated that changes in environmental volatility did not affect the relationship between PEs and LRs (Figure 5B).

Despite being a frequently used analytic approach in the literature, one might argue that block-wise fitting of these models would create a mismatch between the learning strategy employed by the participants versus the models, such that models would be informed about the structure of the task ahead of time. Because, to the best of our knowledge, the blocks were administered continuously by Behrens and colleagues, a continuous fitting approach would ensure that learning models are also agnostic to the information about the task structure. Furthermore, the block-wise fitting approach implemented in the preceding section may also work against the strengths of learning models that can perform dynamic learning rate update. We predicted that a continuous fitting approach would compromise the RW model to a greater extent, as it cannot perform a dynamic learning rate update to adapt to changing outcome schedules between stable and volatile blocks. However, this was not supported by strong evidence, as models remained highly comparable based on log likelihood values ($F(3, 71) = .012$, $p = .998$). Other model goodness of fit measures such as sum of Akaike's Information Criterion (AIC) favoured the cubic model which had in total 3 free parameters (i.e. LR, power utility, inverse temperature), by a margin of 0.31 units to the runner-up model which was hybrid Pearce-Hall which had in total 4 free parameters. Bayes Omnibus risk calculated for these comparisons remained identical to fitting the models individually to each block, as stated above.

Using a continuous fit approach, it is also possible to [re-]investigate the volatility effect on human learning rates by taking the summary/mean value of trialwise learning rates that fall within stable versus volatile blocks (note that this approach only applies to models that can generate trialwise learning rates, namely cubic, exponential-logarithmic and hybrid Pearce-Hall models). This approach did not identify any significant volatility effect ($F(1, 51) = .01$, $p = .919$) or volatility by model interaction ($F(2, 51) = .172$, $p = .842$), but a significant main effect of model influencing the exact parameter estimates (($F(2, 51) = 185.56$, $p<.001$). As previously shown in Figure 5A, the estimates of the Pearce-Hall model were substantially higher than those of the other models. It is possible that the lack of volatility effect using this approach may also be due to an imbalance of trial numbers in the original design of Behrens et al, i.e. 120 trials in the stable block versus 170 trials in the volatile block.

Taken together, these results suggest that the models that we newly proposed behave reasonably comparably to the existing RL models such as RW and hybrid Pearce-Hall, and in the context of learning probabilistic reward associations that deliver variable reward amounts their performance cannot be reliably differentiated.

Exponential-logarithmic model provides greater flexibility capturing behavioural adaptations

In order to tease apart the differences between models that can estimate trialwise learning rates further, we analysed another set of previously published data[40]. 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 somewhat 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/environmental noise).

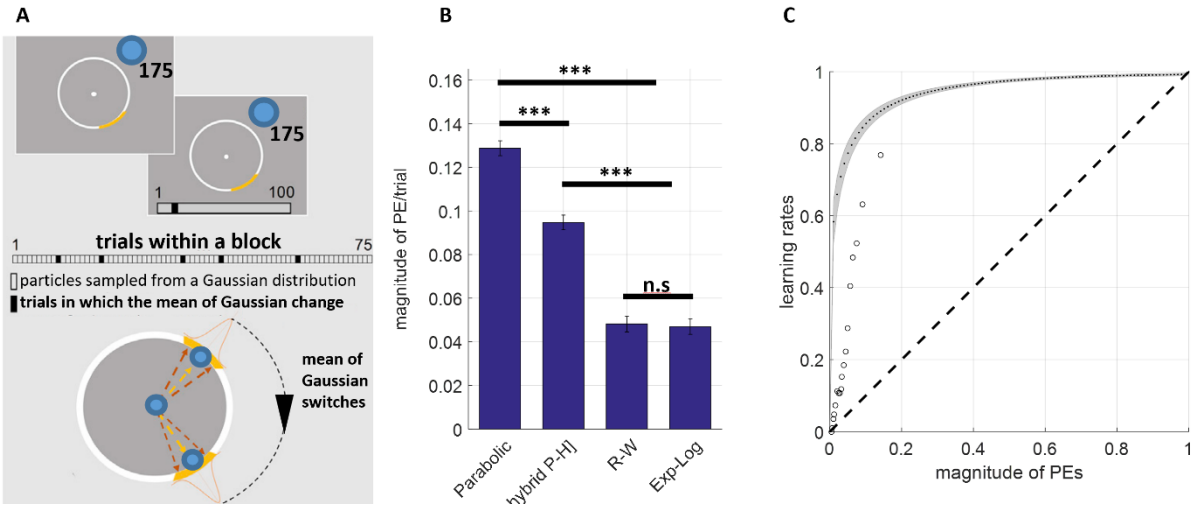


Figure 6. Learning rate results from the reanalysis of the bucket task. (A) Participants try to catch blue particles initiating from the centre of a circular grid with a virtual yellow bucket in order to bank as much points as possible. The particle lending trajectories are drawn from a Gaussian distribution 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, among models that can estimate trialwise learning rates, 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].***p<.001 **(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 reported by Vaghi and colleagues indicated by circular markers, curve with grey shading denotes the trajectory estimated by the exponential-logarithmic model.

In this scenario, we show that the exponential-logarithmic model can account for participant behaviour significantly better than the cubic model ($t(48) = 14.168$, $p<.001$) and the hybrid Pearce-Hall model ($t(48)= 9.514$, $p<.001$). The exponential logarithmic model behaves comparably with the Rescorla-Wagner model ($p = .807$). However, as highlighted before, the R-W model would not allow

447 investigation of the relationship between learning rates and prediction errors. The reason why
448 exponential-logarithmic model performs better than the cubic model becomes evident when we plot
449 the trajectory of the estimated LRs over the range of the prediction errors— participants utilise
450 increasingly higher LRs which gradually plateau when the magnitude of prediction errors exceeds
451 0.2, also revealing a non-parabolic trajectory. These results are widely in agreement with the initial
452 model-free (i.e. how much participants updated their bucket positions from the previous trial) LRs
453 reported by Vaghi and colleagues (2017, circular data points shown in Figure 6C).

454 ***Physiological correlates and temporal properties of learning rates***

455 We also conducted our own learning experiment in an independent cohort to investigate the
456 physiological correlates and temporal properties of LRs versus prediction errors in a task
457 environment with volatility and noise. Thirty-two healthy volunteers (see Supplementary Table 1 for
458 demographic features) completed a reward magnitude learning task which was adapted and
459 simplified from an affective learning task that served as a control experiment reported in one of our
460 earlier studies[16]. On each trial, participants were asked to choose between two abstract shapes and
461 obtained rewards. The reward associated with the chosen shape was represented as a green filling
462 inside a bar which designated 100 available points per trial. This maximum reward amount was
463 distributed between the options, such that if one shape had more reward points (green filling), the
464 other one would have less (i.e. empty grey area within the bar, Figure 7). Participants completed 4
465 blocks of 60 trials each, covering a 2 x 2 factorial design of high/low volatility and high/low noise,
466 making the task environment similar to the simulation environment shown in Figure 4. Blocks were
467 separated by a brief resting period and participants were informed that the outcome schedules
468 might change in each block. Each block also contained novel shape pairs, visually distinct from the
469 shapes that participants already learnt about. In this experiment, because the participants took
470 breaks between distinct task blocks, fitting the RW model separately to each task block can be
471 deemed less problematic.

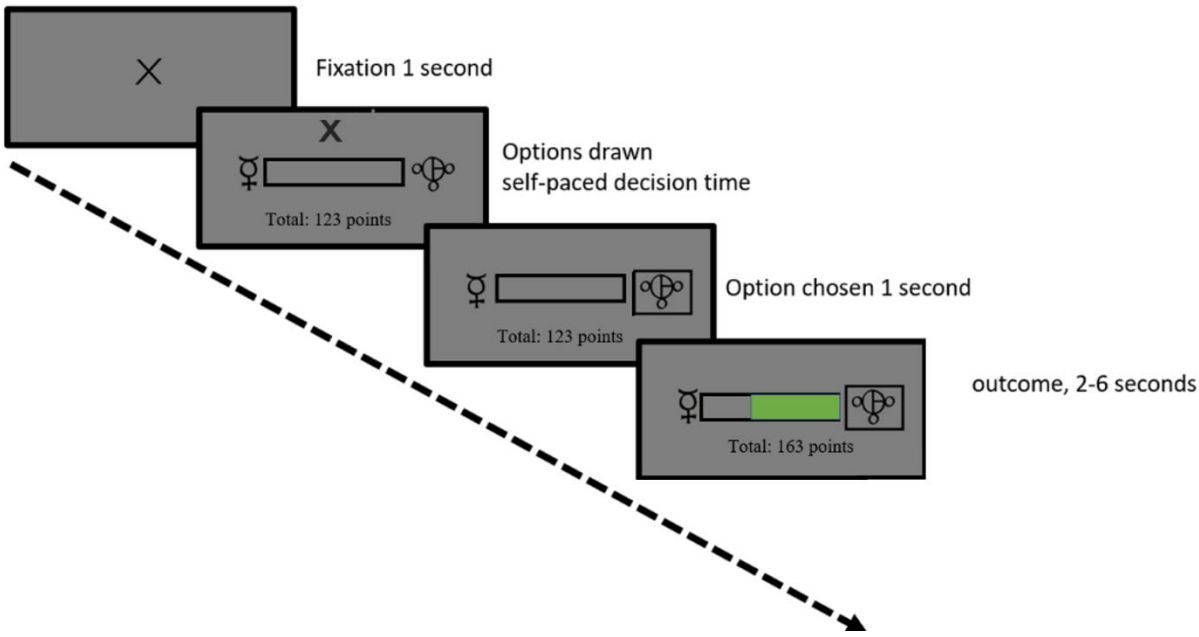


Figure 7. The timeline of the reward magnitude learning task. Following a 1 second fixation period, participants had to choose using the left and right arrow keys between two abstract shapes. They were asked to choose the shape that they thought would give them a larger amount of reward and then instructed 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 6A, also see Pulcu & Browning, 2023 for optimal learning rate adjustment performed by a Bayesian Ideal Observer model in task environments where volatility and noise are independently manipulated[32]). 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 ($F(1, 29) = 5.777$, $p=.023$). Higher environmental noise reduced LRs in volatile environments (line with information theoretic predictions[25], $F(1,29)=7.496$, $p=.01$) but did not influence LRs in stable environments (Supplementary Figure 6A). As expected, there was a main effect of the type of block participants started the task off with in this noise effect ($F(1,29)=4.237$, $p=.049$). Participants' self-reported judgements on environmental volatility and noise, collected at the end of each task block, did not reflect the differences between the conditions, and consequently did not reveal any significant main effects or interaction terms (Supplementary Figure 6B, all $p>.280$). We reanalysed participant choice behaviour using the exponential-logarithmic model, considering that it demonstrated comparable behaviour with the RW model in the preceding sections, and it can be fully reduced to the RW model as shown in Figure 3 [and legends] making these models overall easier to compare. The model log likelihoods accounting for how well they fit to participant choice behaviour were not significantly different between the exponential-logarithmic model, RW with 4

LRs (i.e. a separate learning rate for each block) and RW with one learning rate fitted to all blocks all at once (F(2,95)=2.28, p=.108). Other model selection metrics such as Bayesian exceedance probabilities would favour exponential-logarithmic model among all models fitted to all task blocks all at once (xp= 0.47), whereas group -wise sum of AIC values favoured models that cannot estimate trialwise learning rates, i.e. RW model with 1 (AIC: 7667) or 4 learning rates (AIC: 7199) over the exponential logarithmic model (AIC: 7727). However, the fitting of the hybrid Pearce-Hall was worst (AIC: 8087). Bayesian Omnibus Risk was again calculated to be high at 0.9648.

The relationship between LRs and prediction errors as estimated by the exponential-logarithmic model is shown in Figure 8A (for each subject in our sample, estimated and plotted individually), and reveals a mixture of trajectories: some resembling those shown for the data from Vaghi et al., in the preceding section, others similar to the parabolic trajectory shown in the re-analysis of data from Behrens et al., and also RW-like constant with a subtle and gradual linear increase over the PEs. This pattern of LR/PE relationship was mostly expected as the task design had elements from both of these previously published studies. The LR trajectories further highlight that exponential-logarithmic model can flexibly capture individual variability in RL. The PEs generated under this model demonstrates that in environments with variable [as opposed to binary] outcomes, prediction errors encode the structure of the environment (Supplementary Figure 5).

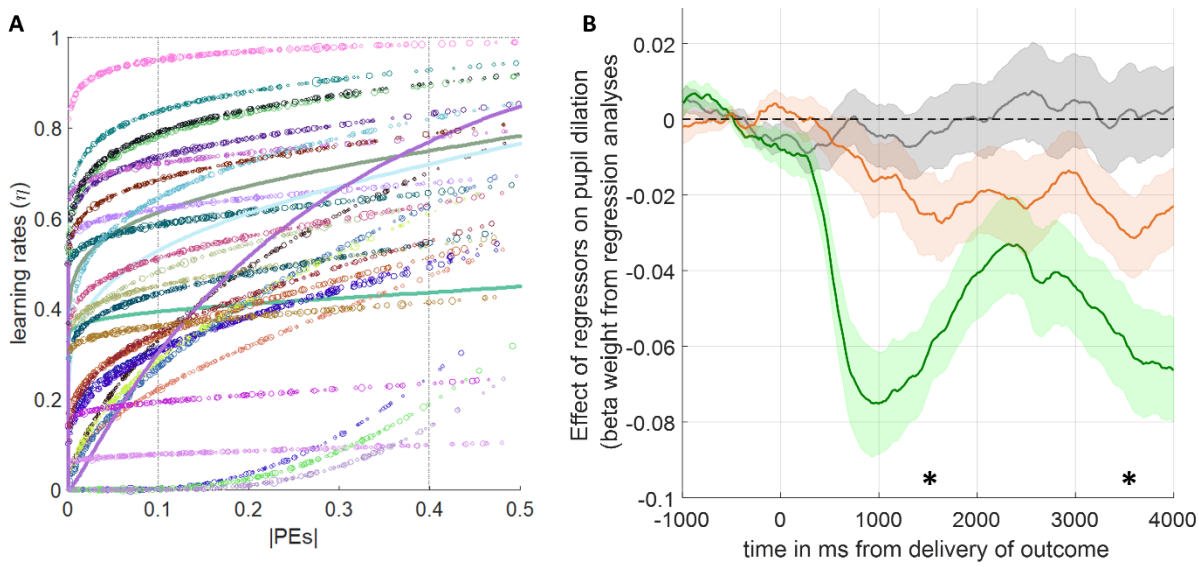


Figure 8. The LRs and their pupillary correlates. (A) Relationship between the absolute value of prediction errors and LRs. The maximum value on x-axis capped at 0.5, as majority of the participants did not experience prediction errors higher than this value. Each single-subject is represented as different 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). Error shading designate ±1 SEM. *p<.05 designate bins of 1000ms in which the average pupil dilation is significantly different relative to baseline.

Subsequently, we investigated the physiological correlates and 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[41], using an ordinary least squares regression model that we also validated in a number of our previous studies[16,32,42,43]. Our prediction was that LRs should be instantiated on the same trial (in line with the cubic and exponential-logarithmic models we introduced earlier, and in contrast to the predictions of the Pearce-Hall model), 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)=-.2589 corresponding to $p<1.57x10^{-27}$; population average r-values significantly lower than zero, p=.0173). Average pupil response encoding LRs was significantly lower than baseline from 1000ms onwards in the outcome delivery period (significance at peak t(27)=-2.969, p=.006). 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 5) did not influence pupil dilation (Figure 8B, grey line). Furthermore, we reanalysed the pupillary data with the identical regression model but using the parameter estimates from the hybrid Pearce-Hall model that can also estimate learning rates on a trial by trial basis, but makes a different temporal prediction about their properties. Learning rates estimated by the Pearce-Hall model did not have a significant physiological correlate (significance at peak t (27) = -1.970, p= .059, Supplementary Figure 7). Therefore, we think pupillary/physiological data lends greater support to the exponential logarithmic model which also had better behavioural model fitting metrics relative to the Pearce-Hall model.

We also report results from iterative stochastic generate-recover simulations which indicated desirable parameter recovery metrics for the exponential-logarithmic model (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[19,40] (in Figures 5, 6 and 8) demonstrated that newly proposed models perform comparably well with the existing RL models while generating reliable 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 4), and pupillary data we collected during human reward-guided learning lend preliminary evidence to support 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) to parse complex higher-order patterns of observed variance in the dependent measure (i.e. the assay of neural activity) into components attributable to a particular source (i.e. experimental or nuisance factors interacting with one another in particular combinations). In modelling functional neuroimaging, GLMs 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. Here, it is worthwhile to highlight that we compared these models in environments where reward (r) distributions, or naturally the reward probabilities, were in the normalised space ($0<r<1$). Future studies may investigate the degree to which these common implementations pose a limitation to the generalisability of these models to real-world scenarios such as consumer behaviour[44].

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[45,46] 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][47], 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 markedly outweigh the lower complexity afforded by the traditional 1-parameter RW model. Although we provided some preliminary 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. Furthermore, it is important to highlight that, analytically, the pupillary data is similar to fMRI data obtained from a single voxel, so future neuroimaging studies using these models can reveal important insights about a wider neural circuitry associated with RL and integration of value information during learning.

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 only the expected uncertainty (i.e. noise) is experimentally manipulated[45,46] 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 6), 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 6). The output of our cubic model (i.e. variable, trial-wise LRs) is similar to a recently described neurophysiological reward-dependent metaplasticity model that can account for the learning behaviour of non-human primates[48]. 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[24]. Needless to highlight, the cubic model is still 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. Particularly, in RL tasks in which participants are expected to learn hidden probabilities such as the one used by Behrens and colleagues, the cubic model's a priori assumptions may be reasonably accurate as we showed in Figure 5B.

In the past two and a half decades, cognitive computational neuroscience has 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[49]. One study 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[13]. On the other hand, three key studies that manipulated the volatility (i.e. the unexpected uncertainty) of reward learning environments showed that volatility of the environment, which in return correlates significantly with LRs, are encoded in the dorsal and subgenual anterior cingulate cortex (ACC) in both non-social and social learning environments[5,19,49]. 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[47]. 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.

Finally, we also hope these models can aid understanding behaviour in psychiatric conditions associated with learning impairments. A recent computational meta-analysis indicated that RL impairments in patients with depression and anxiety load on to reward sensitivity rather than

learning rates[50]. However, our current understanding of impairments in psychiatric conditions rely on the assumption that learning rates are constant as often estimated by the RW model. Here, we showed different lines of evidence that this may not be correct and utilising our models to analyse the relationship between LRs and PEs can pave the way to novel insights in mental health conditions that remain currently unknown.

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

692 across participants. The information content of the reward outcomes was varied across the four
693 blocks by altering the volatility and noise associated with the stimulus-outcome pair. The outcomes
694 for each block were thus generated following a 2 x 2 factorial of high/low volatility and high/low
695 noise schedule (Supplementary Figure 4). For the entire task, the quantity and variety of stimuli
696 remained constant in order to minimise disparities in luminance between trials. At the end of the
697 experiment, the outcome of 20 randomly selected trials were converted to real money which was
698 added to a baseline payment of £10. The participants were told to bank as many points as possible
699 at the beginning of the experiment. After each block, participants were asked to rate the number of
700 times they thought the average reward shifted from low to high, and how wide they thought the
701 range of rewards were around the average reward value (Supplementary Figure 6B).

702 **Pupillometry Preprocessing and Analysis.** We followed identical steps to our previous eye tracking
703 studies in preprocessing and analysing pupillary data[16,32,42,43]. Pupillometry data was collected to
704 examine the participants' physiological responses during the reward-guided learning task. Eyeblinks
705 were removed from the data by using the Eyelink system's built-in filter. This was followed by a
706 linear interpolation to compensate for missing data points. The resulting pupillary trace was
707 processed through a low pass Butterworth filter with a cut-off of 3.75 Hz, and then z-transformed
708 across the experimental session[51]. The pupillary correlates of the model-based regressors were
709 assessed by fitting a multiple linear regression model on baseline corrected pupil data. If over half of
710 the data from the outcome period had been interpolated, the affected trial was excluded from the
711 analysis. Participants who had missing data more than 50% of the trials were excluded from the
712 analysis, leading to the exclusion of four cases in total.

713 The pupillary multiple linear regression model had the following regressors: an intercept, trial
714 number (as a proxy for tiredness), chosen reward outcome, PEs and LRs and average number of
715 black pixels displayed on otherwise grey/neutral screen as shown in the task timeline (Figure 7).

716 **Acknowledgements**

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

References

1 Angela, J. Y. & Dayan, P. Uncertainty, neuromodulation, and attention. *Neuron* **46**, 681-692 (2005).

2 Erev, I. & Roth, A. E. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American economic review*, 848-881 (1998).

3 Sutton, R. S. & Barto, A. G. *Introduction to reinforcement learning*. Vol. 135 (MIT press Cambridge, 1998).

4 Rmus, M., Zou, A. & Collins, A. G. Choice type impacts human reinforcement learning. *Journal of Cognitive Neuroscience* **35**, 314-330 (2023).

5 Behrens, T. E., Hunt, L. T., Woolrich, M. W. & Rushworth, M. F. Associative learning of social value. *Nature* **456**, 245-249 (2008).

6 Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593-1599 (1997).

7 O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H. & Dolan, R. J. Temporal difference models and reward-related learning in the human brain. *Neuron* **38**, 329-337 (2003).

8 Franklin, N. T. & Frank, M. J. A cholinergic feedback circuit to regulate striatal population uncertainty and optimize reinforcement learning. *Elife* **4**, e12029 (2015).

9 Khorsand, P. & Soltani, A. Optimal structure of metaplasticity for adaptive learning. *PLoS computational biology* **13**, e1005630 (2017).

10 Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J. & Lamm, C. Using reinforcement learning models in social neuroscience: frameworks, pitfalls and suggestions of best practices. *Social Cognitive and Affective Neuroscience* **15**, 695-707 (2020).

11 Silver, D. *et al.* Mastering the game of go without human knowledge. *Nature* **550**, 354 (2017).

12 Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *nature* **529**, 484 (2016).

13 Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A. & Daw, N. D. Differential roles of human striatum and amygdala in associative learning. *Nature neuroscience* **14**, 1250 (2011).

14 Prelec, D. The probability weighting function. *Econometrica*, 497-527 (1998).

15 Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X. & Bishop, S. J. Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature neuroscience* (2015).

16 Pulcu, E. & Browning, M. Affective bias as a rational response to the statistics of rewards and punishments. *Elife* **6** (2017).

17 Rescorla, R. A. & Wagner, A. R. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory* **2**, 64-99 (1972).

18 Pearce, J. M. & Hall, G. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review* **87**, 532 (1980).

19 Behrens, T. E., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. Learning the value of information in an uncertain world. *Nature neuroscience* **10**, 1214-1221 (2007).

20 Mathys, C., Daunizeau, J., Friston, K. J. & Stephan, K. E. A Bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience* **5**, 39 (2011).

21 Palminteri, S. & Lebreton, M. The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in cognitive sciences* (2022).

22 Louie, K. Asymmetric and adaptive reward coding via normalized reinforcement learning. *PLoS Computational Biology* **18**, e1010350 (2022).

23 Roesch, M. R., Esber, G. R., Li, J., Daw, N. D. & Schoenbaum, G. Surprise! Neural correlates of Pearce–Hall and Rescorla–Wagner coexist within the brain. *European Journal of Neuroscience* **35**, 1190-1200 (2012).
24 Soltani, A. & Izquierdo, A. Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience*, 1 (2019).
25 Pulcu, E. & Browning, M. The misestimation of uncertainty in affective disorders. *Trends in Cognitive Sciences* (2019).
26 Charpentier, B., Senanayake, R., Kochenderfer, M. & Günnemann, S. Disentangling epistemic and aleatoric uncertainty in reinforcement learning. *arXiv preprint arXiv:2206.01558* (2022).
27 Piray, P. & Daw, N. D. A model for learning based on the joint estimation of stochasticity and volatility. *Nature communications* **12**, 6587 (2021).
28 ST, Y. The bit juggler. *nature physics* **19**, 919 (2023).
29 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).
30 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).
31 Piray, P. & Daw, N. D. A simple model for learning in volatile environments. *PLoS computational biology* **16**, e1007963 (2020).
32 Pulcu, E. & Browning, M. Humans adapt rationally to approximate estimates of uncertainty. *bioRxiv*, 2023.2011. 2026.568699 (2023).
33 Pulcu, E. *et al.* Temporal discounting in major depressive disorder. *Psychological Medicine*, 1-10 (2013).
34 Murphy, J., Vuchinich, R. & Simpson, C. Delayed reward and cost discounting. *The Psychological Record* **51**, 571-588 (2001).
35 Kahneman, D. & Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263-291 (1979).
36 Stott, H. P. Cumulative prospect theory's functional menagerie. *Journal of Risk and uncertainty* **32**, 101-130 (2006).
37 Daw, N. D. Trial-by-trial data analysis using computational models. *Decision making, affect, and learning: Attention and performance XXIII* **23**, 1 (2011).
38 Luce, R. D. The choice axiom after twenty years. *Journal of mathematical psychology* **15**, 215-233 (1977).
39 Wang, J. X. *et al.* Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience* **21**, 860 (2018).
40 Vaghi, M. M. *et al.* Compulsivity reveals a novel dissociation between action and confidence. *Neuron* **96**, 348-354. e344 (2017).
41 Behrens, T. E., Hunt, L. T. & Rushworth, M. F. The computation of social behavior. *science* **324**, 1160-1164 (2009).
42 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).
43 Pike, A. C. *et al.* Adaptive learning from outcome contingencies in eating-disorder risk groups. *Translational psychiatry* **13**, 340 (2023).
44 Schulz, E. *et al.* Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences* **116**, 13903-13908 (2019).
45 Diederen, K. M. & Schultz, W. Scaling prediction errors to reward variability benefits error-driven learning in humans. *Journal of Neurophysiology* **114**, 1628-1640 (2015).

46 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).

47 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).

48 Farashahi, S. *et al.* Metaplasticity as a neural substrate for adaptive learning and choice under uncertainty. *Neuron* **94**, 401-414. e406 (2017).

49 Vinckier, F. *et al.* Confidence and psychosis: a neuro-computational account of contingency learning disruption by NMDA blockade. *Molecular Psychiatry* **21**, 946-955 (2016).

50 Pike, A. C. & Robinson, O. J. Reinforcement Learning in Patients With Mood and Anxiety Disorders vs Control Individuals: A Systematic Review and Meta-analysis. *JAMA psychiatry* (2022).

51 Nassar, M. R. *et al.* Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature neuroscience* **15**, 1040-1046 (2012).

52 Nassar, M. R. *et al.* Age differences in learning emerge from an insufficient representation of uncertainty in older adults. *Nature Communications* **7**, 11609 (2016).

855 ## Supplementary Tables and Figures

856 **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)

857

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

861

862

863

864

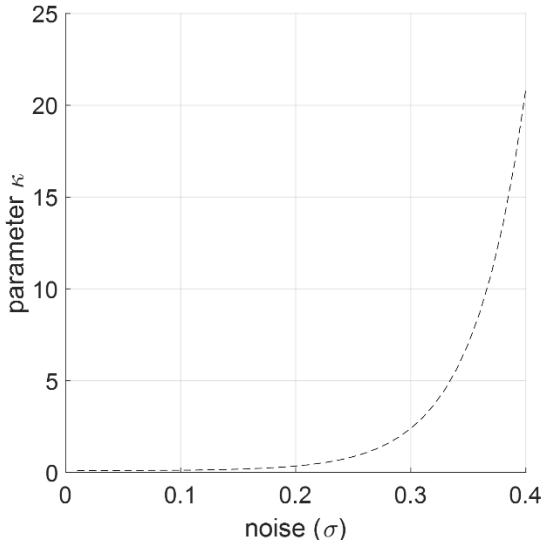
865

866

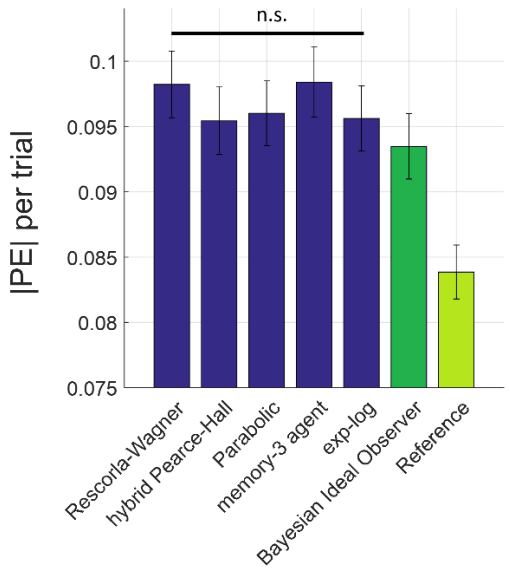
867

868

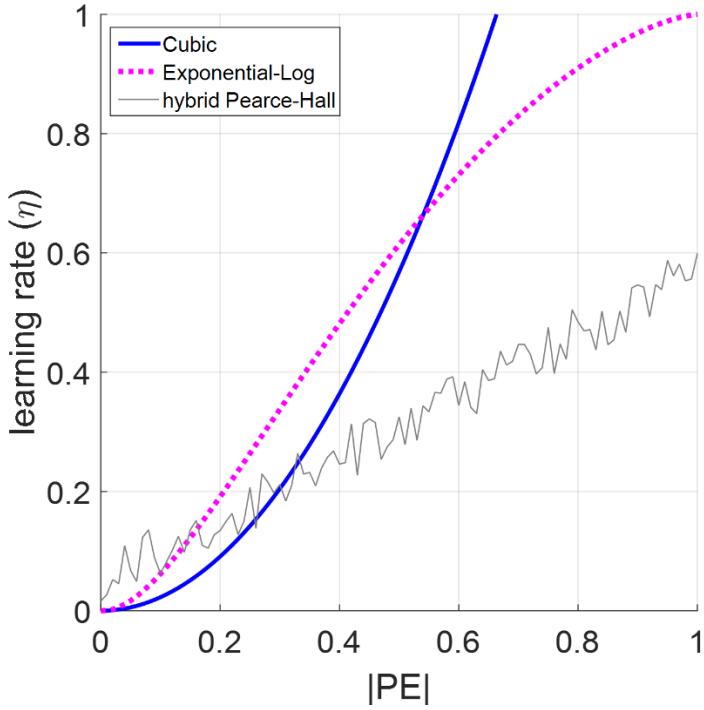
869



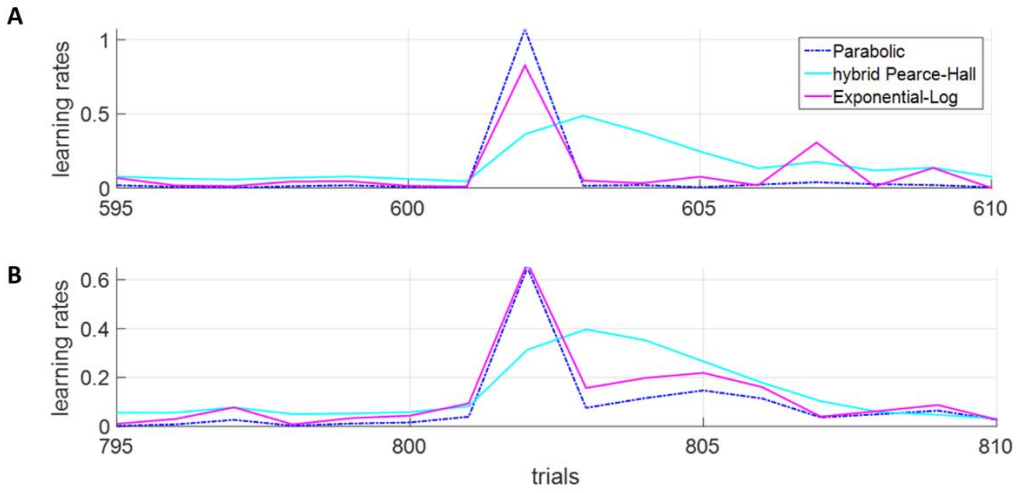
Supplementary Figure 1. The relationship between environmental noise and parameter κ which underlies Figure2.



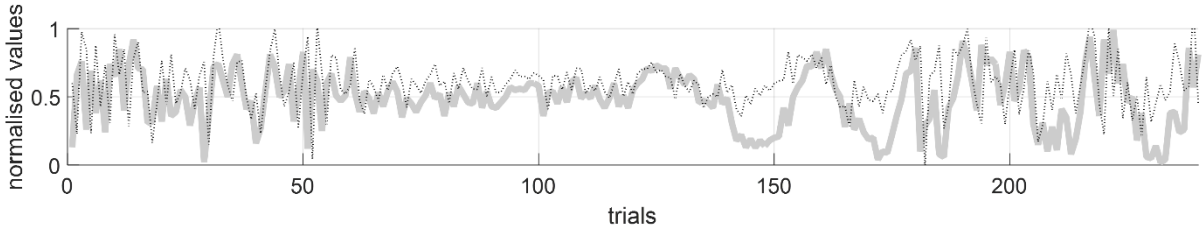
Supplementary Figure 2. The average magnitude of prediction errors (|PE| on y-axis) under different learning models. Across 1500 trials in the simulated task environment (Figure 4), the RL models lead to average PEs which were not statistically significantly different. Based on an exponential relationship 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[25,29,32] 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).



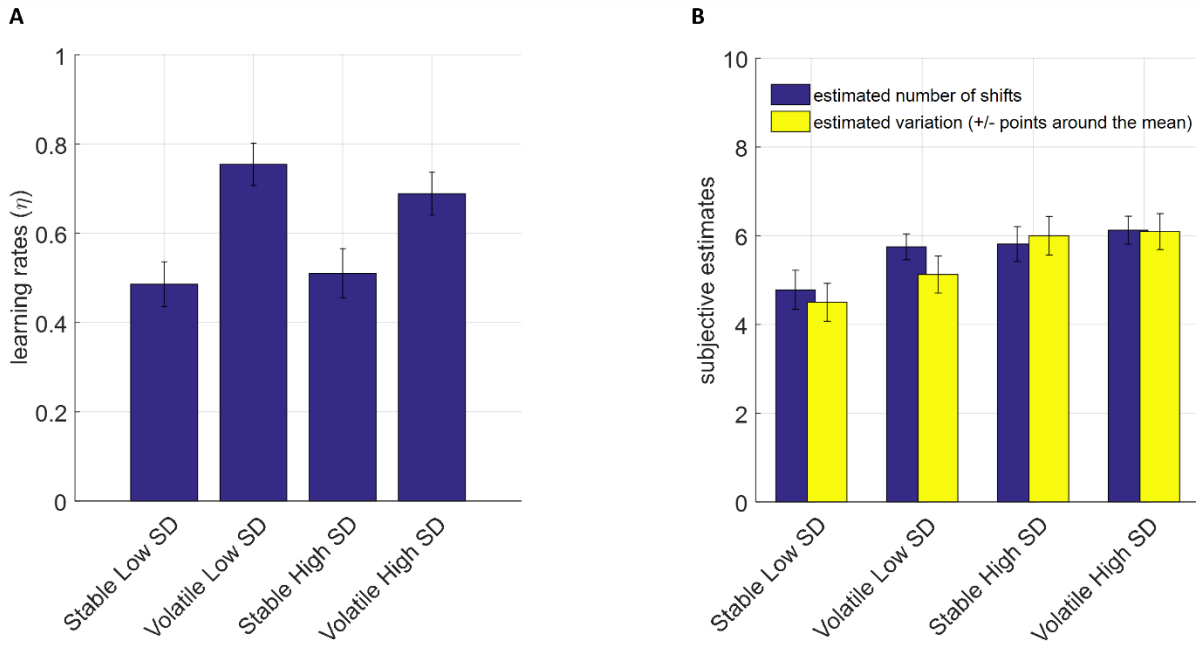
Supplementary Figure 3. A nonlinear relationship between the absolute magnitude of prediction errors (|PE|) and LRs. The exponential-logarithmic model which behaves marginally, but not statistically, better in the simulation environment (Figure 4) have steeper LRs relative to the cubic model for observations 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. The hybrid Pearce-Hall model would suggest a stochastic linear relationship between |PE|and learning rates, as the learning rate on a given trial is also a function of the learning rate from one previous trial. The value of the ω parameter, that is 0.48, suggests that the learning rates would be influenced by the magnitude of the prediction errors to a greater extent leading to the observed linear relationship (Eq.2).



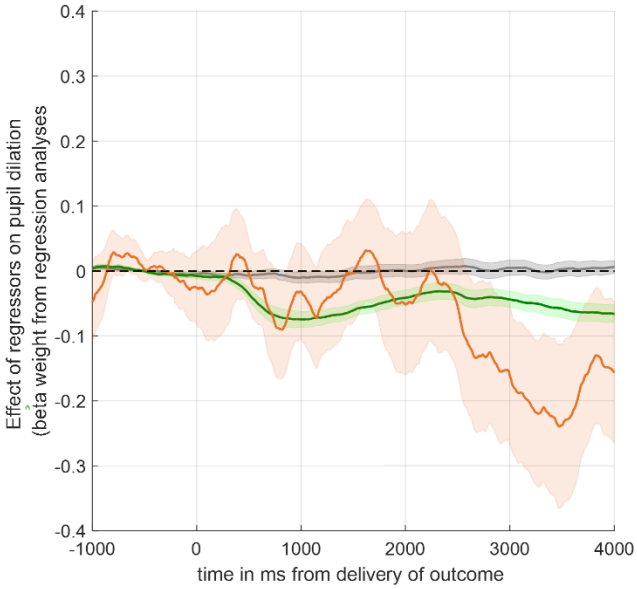
Supplementary Figure 4. Learning rate acceleration between hybrid Pearce-Hall, cubic and exponential-logarithmic models. The hybrid Pearce-Hall model is slower to accelerate and decelerate LRs. The trial numbers on x-axes refer to the task environment shown in Figure 4. Examples are shown for **(A)** 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[51,52].



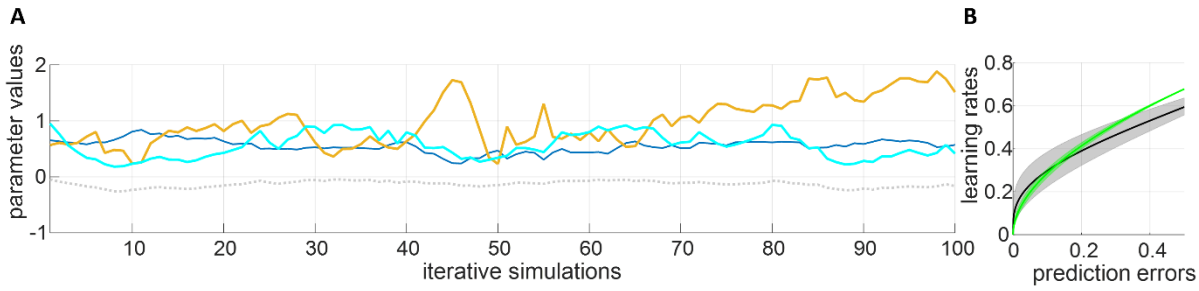
Supplementary Figure 5. Task environment of the reward magnitude learning task (thick grey line) as experienced by a participant. Block orders were randomised. Prediction errors from the exponential-logarithmic model superimposed (thin dashed black line), demonstrating that prediction errors globally encode the structure of the environment.



Supplementary Figure 6. Behavioural results. (A) LRs measured by the RW model fitted to each block separately, demonstrate an overall normative behaviour in human learners[32]. **(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 7. Physiological correlates of learning rates (red line with corresponding error shading) under the Pearce-Hall model. Error shading designate ±1 SEM.



Supplementary Figure 8. 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 (thin navy line). Orange and cyan lines designate the gamma and delta parameters of the exponential-logarithmic model. Data stochastically generated by the exponential-logarithmic model is better explained by this model rather than the RW model (means of predictive accuracy based on stochastically generated choices and out of 240 trials in the original task environment: 88.4% vs 80.4%, $t(31)=20.478$, $p<.001$). **(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 the main body of the article (Figure 8A).