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

Erdem Pulcu

University of Oxford

Department of Psychiatry

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:

We are living in a dynamic world in which stochastic relationships between cues and outcome events create different sources of uncertainty[1] (e.g. the fact that 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 which is very important 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[4-6], as well as machine learning[7,8]. This paper demonstrates the suitability of novel update rules for RL, based on a nonlinear relationship between prediction errors (i.e. difference between the agent's expectation and the actual outcome) and learning rates (i.e. a coefficient with which agents update their beliefs about the environment), that can account for learning-based adaptations in the face of environmental uncertainty. These models illustrate how learners can flexibly adapt to dynamically changing environments.

Keywords: reinforcement learning, uncertainty, learning rates, prediction errors, computational modelling

Introduction

A number of previous studies showed that human learners adjust their learning rates to the information content of outcomes[9,10], utilising a higher learning rate for outcomes more informative for predicting the future state of the environment. In computational neuroscience, learning rates are numerically estimated from participant choices by using different reinforcement learning (RL) algorithms[11-13], which are in effect approximations of the Bayesian models[14,15] that can also estimate otherwise hidden generative rules of dynamically changing environments. The Rescorla-Wagner (RW) algorithm is probably the most widely used RL model[11] 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. 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 learning rates during that period. This is an approach that uses the RW model as a tool of measurement, implemented by a number of key RL studies[4,9,10,14] consistently showing that human learning rates 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 demonstrate that learning rates are malleable. 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[12,13,16] aim to mitigate this limitation by updating learning rates on a trial-by-trial basis. For example, one common implementation is based on a weighted integration of the learning rates from previous trials with the magnitude of the absolute value of the prediction errors (PE; i.e. the difference between the agent's expectation and the real outcome):

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

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. 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:

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

where λ is another free-parameter used for smoothing the transition of the probability update.

Assumptions underlying the novel update rules

Different sources of uncertainty dictates what kind of learning behaviour would be optimal in a given environment[1,17,18]. *Expected uncertainty* relates to intrinsic variability in the occurrence or the nature of an event over and above that can 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, as 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 with a relatively low learning rate. However, there can be times when learned associations change over time. In this case, 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*"[14]). The effect of unexpected uncertainty is to reduce how informative 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 learn more rapidly over a shorter period of time, 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 learning rates 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 learning rates, the exact nature of this relationship has not been explored yet. Below, the manuscript evaluates the suitability of two novel update rules for RL that establish a nonlinear relationship between learning rates and the magnitude of the PEs.

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 the PEs can be is a vector of -1s. Another widely established assumption is that the agents' learning rates should be 0 if the PE is 0. 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 nonlinearity in wide majority of cognitive processes (e.g. risk perception[19], stochastic choice[20,21] etc.) in human and non-human primates, a simple functional form satisfying both of these assumptions establishes a parabolic relationship between the PEs and learning rates:

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

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

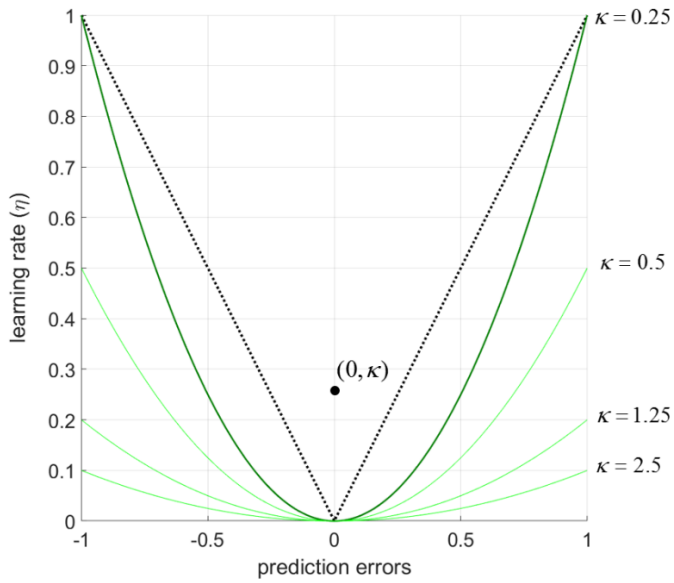


Figure 1. The relationship between prediction errors and learning rates in the Dynamic Adjustment model. In environments with high unexpected but low expected uncertainty, one would assume a somewhat linear relationship between prediction errors and learning rates (in a V shape designated by the black 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), with increasing values of κ flattening the relationship between prediction errors and learning rates, leading to lower learning rates 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 allow prediction errors to be converted to learning rates even more steeply.

The formula of the novel update algorithm (i.e. updating PEs with corresponding learning rates on each trial) for RL is then written as:

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

The free-parameter κ accounts for individual variability in how steeply PEs accelerate learning rates during RL. 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 learning rates to the PEs would not be an ideal way to learn from the environment. This is because, in 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 can also be interpreted as 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[17]. A recent analysis of normative learning models demonstrated that learning rates decay exponentially in noisy environments with high expected uncertainty[22]. 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 learning rates with respect to the magnitude of prediction errors at a given level of environmental noise. Under these assumptions nonlinearity of learning rates guiding behavioural adaptations is shown in Figure 2. The distribution of these learning rates is globally in line with results reported by other research groups which showed a linear-stochastic relationship between learning rates estimated from participant choices and prediction errors as a function of expected uncertainty/environmental noise[23].

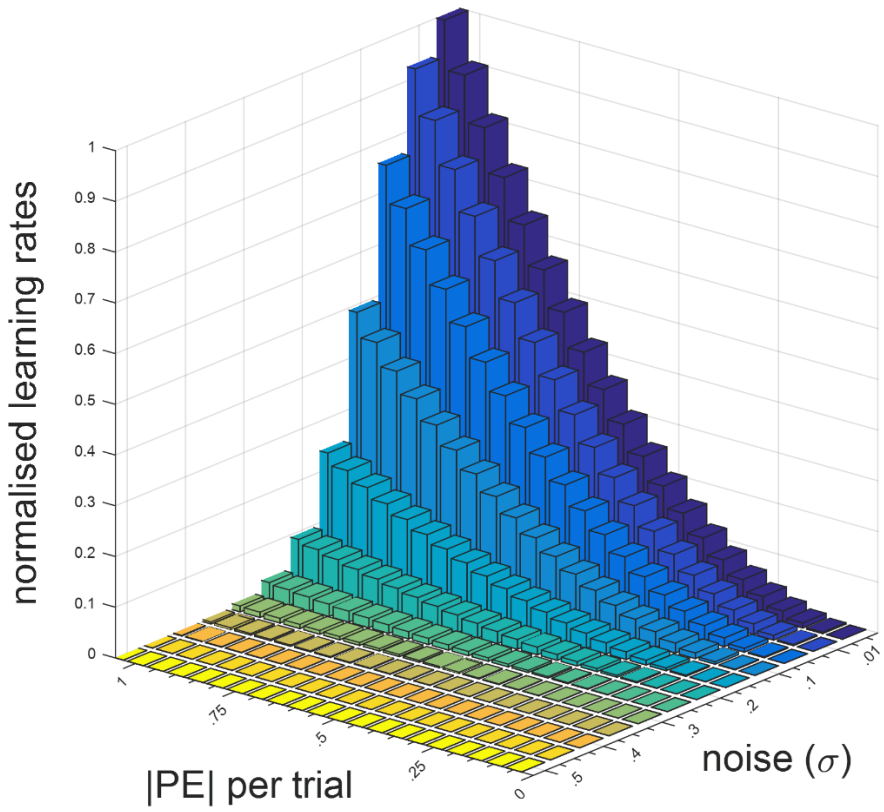


Figure 2. Nonlinearity in trialwise learning rates as a function of prediction errors and environmental noise. Increasing levels of expected uncertainty (σ) diminishes the relationship between PEs and learning rates. In these situations the agents should no longer attribute high magnitude prediction 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 learning rates 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. ϒ denotes the Euler-Mascheroni constant.

Another approach with which one can extract trialwise learning rates is by exploiting the mathematical properties of a 2-parameter [probability] weighting function[24]. This approach would assume an exponential-logarithmic relationship between PEs and learning rates and allows even more flexibility than the parabolic/cubic update rule described above, as it is free of any *a priori* assumptions about the shape of the relationship between PEs and learning rates:

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

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 learning rates are parabolic (Eq. 5).

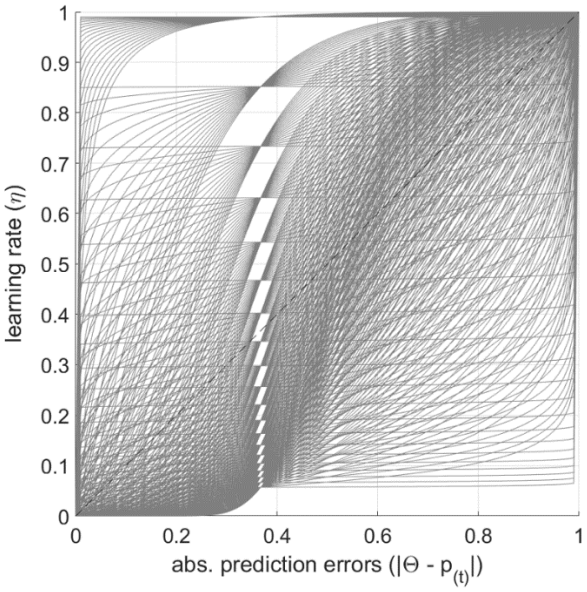


Figure 3. Nonlinearity in trialwise learning rates as a function of prediction errors under an exponential-logarithmic assumption. An exponential-logarithmic weighting function allows greater flexibility in capturing individual variability in how PEs shape learning rates. The parameters δ and λ are coarsely 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 common Rescorla-Wagner learning rule.

Results

Simulated data. We simulated the behaviour of these RL models (i.e. Rescorla-Wagner, hybrid Pearce-Hall, parabolic 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). Learning rates were estimated under each RL model by minimising the average PE per trial throughout the task environment (t=1500 trials).

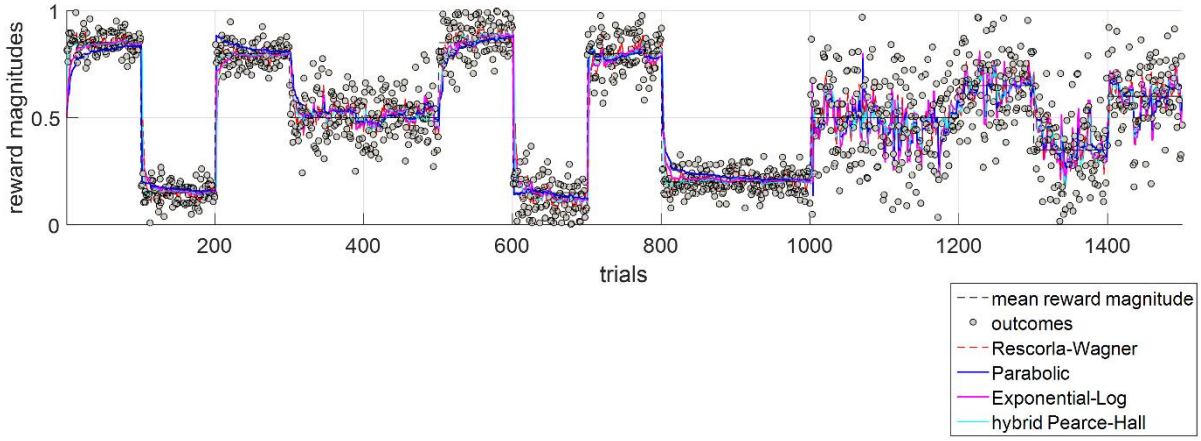


Figure 4. The behaviour of different reinforcement learning models in a task environment in which unexpected and expected uncertainties were independently manipulated. On each trial the task environment returned a reward outcome between 0 and 1 (y-axis). The outcomes were generated to be variable from 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 trial 500, 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-.20), 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 parabolic 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 significantly different ($F(4, 7495) = 0.32$, $p = 0.863$, Supplementary Figure 1), meaning that all models can account for behavioural adjustments comparably well. The nonlinear relationship between PEs and learning rates under the parabolic and the exponential-logarithmic models are illustrated in Supplementary Figure 2.

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

higher than the ones estimated under the parabolic (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 were highly significantly correlated (all $r(1498)>.68$, $p<.0001$). Taken together, these results suggest that the novel parabolic and exponential-logarithmic models can be viable alternatives to the existing RL models, with added benefits of: (i) extracting trialwise learning rates, and (ii) help scrutinising the relationship between PEs and learning rates at single subject and population levels. It is worthwhile to emphasise that the parabolic model is a simpler implementation as it utilises a single free-parameter to extract trialwise learning rates.

Human behaviour. We also report the behaviour of these RL models on a previously published data set in which human participants (N=18) completed a learning task with stable (120 trials) or volatile (170 trials) reward contingencies in different task blocks[14]. In the original work by Behrens and colleagues (2007), 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 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 models) with their corresponding reward magnitudes associated with each option (modulated by a single power utility parameter accounting for participants' risk attitude[19]) and calculating the expected value difference between two options and choosing between them stochastically following the choice probabilities generated by a sigmoid function[20].

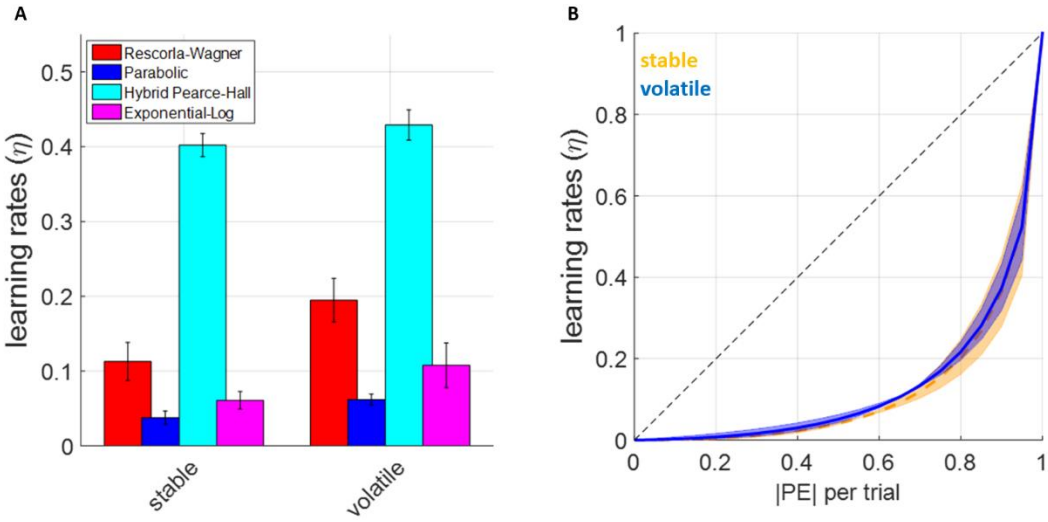


Figure 5. The learning rates estimated under each reinforcement learning model. (A) Average learning rates estimated under each model. Hybrid Pearce-Hall model estimates higher learning rates relative to the Rescorla-Wagner model, whereas parabolic and exponential-logarithmic models estimate overall lower learning rates. **(B)** The relationship between PEs and learning rates estimated by the exponential-logarithmic model shows nonlinearity very similar to the parabolic model, confirming this assumption. Volatility of different task blocks did not lead to significant differences in learning rate trajectories.

Overall, all RL models estimated higher learning rates in the volatile relative to the stable block, replicating the findings previously reported by Behrens et. al., 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 ($F(3,68)=.002$, $p=1$). However, the difference in learning rates between stable and volatile blocks were only statistically significant under the RW and parabolic models (t (34) values 2.114 and 2.079, p values 0.042 and 0.045, respectively). The hybrid Pearce-Hall and the exponential-logarithmic models did not capture a statistically significant volatility effect on learning rates (t (34) =1.057 and 1.460, p=.298 and .153, respectively).

Discussion

The current work describes two novel update rules for human reinforcement learning (RL), both establishing a nonlinear relationship between prediction errors (PEs) and learning rates (Supplementary Figure 2). Both simulated data (Figure 4) and analysis of human behaviour (Figure 5) suggest that the newly proposed models perform comparably with the existing Rescorla-Wagner and Pearce-Hall models. Both of these novel models are able to generate trialwise learning rate estimates from participant choice behaviour. This feature of the models is also shared by the hybrid Pearce-Hall model (Eqs. 2-3) and subsequent numerical analyses show that all models generate learning rates which are comparably decorrelated with the PEs. The main difference between these models is how quickly they would adjust their behaviours in response to environmental change. Under the parabolic and exponential-logarithmic models, learning rates accelerate as soon as there is a dramatic environmental change leading to large PEs, and learning rates subsequently decay as the magnitude of PEs get smaller. Under the hybrid Pearce-Hall model acceleration and deceleration of learning rates happen relatively more gradually as learning rates are weighted integration of learning rates and the PEs from the previous trials (Supplementary Figure 3), determined by a weighting constant (Eq.2) estimated from participant choices. Although all of these models describe an efficient way of interacting with the environment (i.e. accelerating learning rates only in cases in which the changes in the environment are informative), they make different predictions about the temporal properties of learning rates (Supplementary Figure 3). In the hybrid Pearce-Hall model the information is fed forward as the learning rate is a secondary process updated in parallel (Eqs. 2-3), in addition to estimated reward rate, whereas parabolic and exponential-logarithmic models assume that prediction errors would trigger learning rates instantaneously, during the feedback period of the current trial. These differences can be elucidated by future neurophysiology studies which were instrumental in revealing temporal properties of prediction errors[5].

In tasks in which expected uncertainty is experimentally manipulated[25,26], it might be more accurate to rely on the parabolic model (Figure 2) and estimate σ 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. Analysis of human learning behaviour using the exponential-logarithmic model, which is free of any assumptions regarding the shape of the nonlinearity (Figure 3), also indicated a relationship resembling the parabolic form (Figure 5B), providing further evidence in favour of this assumption. Although the output of this parabolic model (i.e. variable, trialwise learning rates) is similar to a recently described neurophysiological reward-dependent metaplasicity model that can account for the learning behaviour of non-human primates[27], under the current 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[17]. Furthermore, the cubic/parabolic model is much easier implement on data collected from human volunteers due to feasibility limitations in electrophysiological recording.

The newly proposed update rules will have important implications for informing neural sciences. Analysis of various functional neuroimaging modalities (e.g. fMRI, MEG, pupillometry, etc.) heavily relies on general linear models (GLMs) which require trialwise estimates of behavioural processes as predictive variables (i.e. regressors). Reliable, trialwise estimates of learning rates is an extremely important issue in neural sciences, considering that the number of studies that investigated brain regions encoding PEs significantly outnumber those which investigated neural regions encoding learning rates. One study using the hybrid Pearce-Hall model described in this paper in a reversal learning task involving aversive outcomes (30% probability of receiving an electroshock) showed that learning rates are encoded in the human amygdala, bilaterally. 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 learning rates, are encoded in the dorsal anterior cingulate cortex (dACC) in both non-social and social learning environments. The converging results presented in this paper demonstrate that the cubic/parabolic learning rule is able to capture a significant main effect of outcome volatility in line with the latter studies, while estimating learning rates on a trial-by-trial basis which are highly significantly correlated with the Pearce-Hall model used in the former study, having the best of both worlds. However, so far no study has independently manipulated the sources of expected and unexpected uncertainties in learning environments (such as the task environment shown in Figure 4), without explicitly informing participants about in which condition they are asked to perform reinforcement learning. Future studies using this full factorial approach can probe where different components of RL models are encoded in the human brain (i.e. striatal regions, bilateral amygdala, dACC) more precisely. These are important directions not only for cognitive/computational neuroscience but also for a better understanding of psychiatric disorders in which learning processes are impaired.

Overall, the cubic/parabolic learning rule described in the current paper can be a valuable alternative to existing RL models as it is simple, intuitive and can extract trialwise learning rates and PEs which are decorrelated with each other. This latter feature of the model can help computational neuroscientists better understand how human brain implements different components of RL models.

Methods

All analyses were implemented on MATLAB 2015b (MathWorks Inc.) and all analyses scripts, including the simulation environment shown in Figure 3, are deposited on the servers of the Open Science Framework.

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

Acknowledgements

The author would like to thank to Prof. Tim Behrens for sharing the behavioural data and a helpful discussion. The author would also like to thank to Prof. Michael Browning and the participants of the Oxford Computational Psychiatry Journal Club for their helpful comments about behavioural modelling.

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

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

6 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).

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

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

9 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).

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

11 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).

12 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).

13 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).

14 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).

15 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).

16 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).

17 Soltani, A. & Izquierdo, A. Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience*, 1 (2019).

18 Pulcu, E. & Browning, M. The misestimation of uncertainty in affective disorders. *Trends in Cognitive Sciences* (2019).

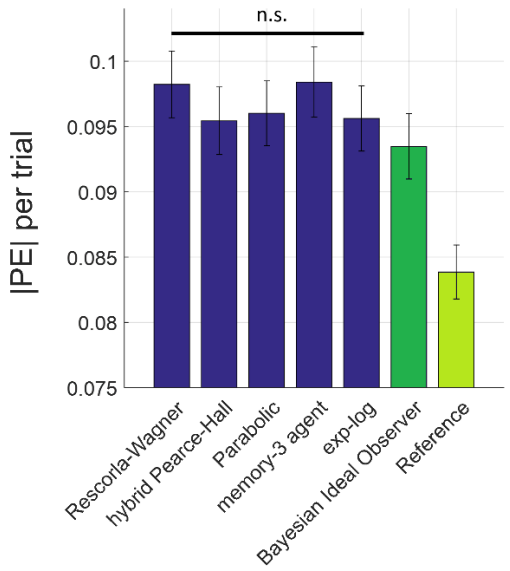
19 Kahneman, D. & Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263-291 (1979).

20 Daw, N. D. Trial-by-trial data analysis using computational models. *Decision making, affect, and learning: Attention and performance XXIII* **23**, 1 (2011).

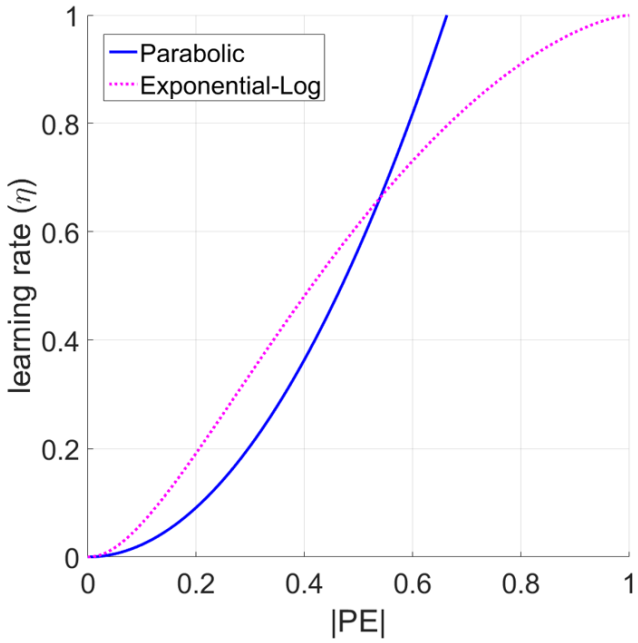
21 Luce, R. D. The choice axiom after twenty years. *Journal of mathematical psychology* **15**, 215-233 (1977).

22 Wang, J. X. *et al.* Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience* **21**, 860 (2018).

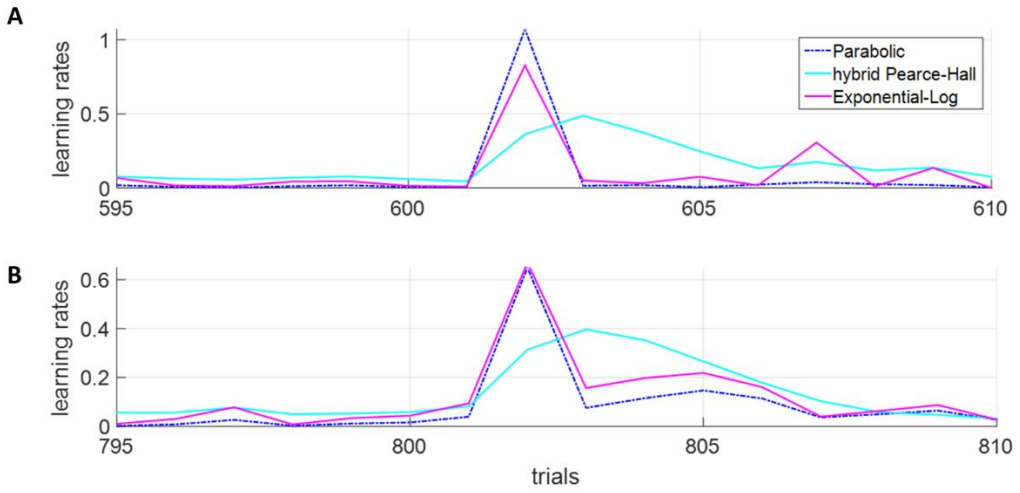
23 Nassar, M. R., Wilson, R. C., Heasly, B. & Gold, J. I. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience* **30**, 12366-12378 (2010).
24 Prelec, D. The probability weighting function. *Econometrica*, 497-527 (1998).
25 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).
26 Diederen, K. M. & Schultz, W. Scaling prediction errors to reward variability benefits error-driven learning in humans. *Journal of Neurophysiology* **114**, 1628-1640 (2015).
27 Farashahi, S. *et al.* Metaplasticity as a neural substrate for adaptive learning and choice under uncertainty. *Neuron* **94**, 401-414. e406 (2017).
28 Nassar, M. R. *et al.* Age differences in learning emerge from an insufficient representation of uncertainty in older adults. *Nature Communications* **7**, 11609 (2016).
29 Nassar, M. R. *et al.* Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature neuroscience* **15**, 1040-1046 (2012).



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[18] 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).



Supplementary Figure 2. A nonlinear relationship between the absolute magnitude of prediction errors (|PE|) and learning rates. The exponential-logarithmic model which behaves marginally, but not statistically, better in the simulation environment (Figure 4) have steeper learning rates relative to the parabolic 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.



Supplementary Figure 3. Learning rate acceleration between hybrid Pearce-Hall, parabolic and exponential-logarithmic models. The hybrid Pearce-Hall model is slower to accelerate and decelerate learning rates. 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) learning rates can marginally be higher than 1[28,29].