Uncertainty-dependent learning bias in value-based decision making

Kitti Bán[1,2], Eszter Tóth-Fáber[3,4], Martin Lages[1,*], Andrea Kóbor[2,*]

1) School of Neuroscience and Psychology, University of Glasgow, Glasgow, United Kingdom
2) Brain Imaging Centre, HUN-REN Research Centre for Natural Sciences, Budapest, Hungary
3) Institute of Psychology, ELTE Eötvös Loránd University, Budapest, Hungary
4) Brain, Memory and Language Research Group, Institute of Cognitive Neuroscience and Psychology, HUN-REN Research Centre for Natural Sciences, Budapest, Hungary
*: Shared senior authorship: these authors contributed equally to this work.
Correspondence: Kitti Ban, ban.kitti@ttk.hu

Abstract

Do we preferentially learn from positive rather than negative decision outcomes? Previous studies indicated that such bias characterises learning during simple reward learning tasks. However, no research has yet confirmed whether learning bias is also present during sequential decision making under uncertainty. To fill this gap, we utilised a complex yet ecologically valid paradigm, the Balloon Analogue Risk Task (BART), which measures risk-taking propensity under uncertainty in everyday decision making. Comparing learning from positive and negative outcomes in the BART has been made possible by the Scaled Target Learning model, which characterises both risk-taking propensity and sensitivity to wins and losses. For the first time, we applied this model to a modified BART paradigm with different levels of perceived uncertainty. Crucially, our analyses revealed learning bias during high levels of uncertainty, under which condition bias was negatively tied to task performance. Furthermore, increased sensitivity to wins compared to losses was linked to more risk-seeking behaviour across all conditions, suggesting that learning bias could mediate risky behaviour. Overall, our results contribute to a more accurate characterisation of reward learning behaviour and suggest that learning bias arises when the level of perceived uncertainty surges.

Key words: Balloon Analogue Risk Task; Bayesian modelling; reinforcement learning; reward; risk

Introduction

Despite abundant evidence exists in support of differential learning from positive and negative decision outcomes, the exact behavioural processes linked to such differential learning, including the potential modulatory role of uncertainty, are yet unclear.

Optimism bias, whereby people overestimate the probability of positive future events and underestimate the probability of negative future events, has been demonstrated in different walks of life (Kuzmanovic & Rigoux, 2017; Sharot et al., 2011; 2012; Shepherd et al., 2013; Weinstein, 1980) and is considered to originate from an asymmetry in belief updating. Evidence from simple instrumental learning tasks with (den Ouden et al., 2013; Palminteri et al., 2017) and without reversals (Frank et al., 2007; Lefebvre et al., 2017; Niv et al., 2012) demonstrated that this asymmetry also characterises basic reinforcement learning, as indicated by higher positive compared to negative learning rates. Additionally, a recent study by Palminteri (2023) revealed learning bias to be present in 9 different two-armed bandit tasks with binary probabilistic outcomes and feedback, further suggesting that learning bias is a universal phenomenon in reinforcement learning. Harada (2020) also revealed such a learning bias in the Iowa Gambling Task, a more complex and ecologically valid paradigm compared to the more abstract instrumental learning tasks used in the above studies. However, this bias disappeared with the introduction of dynamic, trial-wise learning rates in their reinforcement learning model (Harada, 2020).

To our knowledge, no studies yet reported whether this positivity bias in learning also exists during sequential decision making under uncertainty. To fill this gap, we utilised the Balloon Analogue Risk Task (BART; Lejuez et al., 2002), a popular and intuitive paradigm that measures risk-taking propensity by emulating an uncertain decision context with probabilistic rewards. In each experimental trial, participants are presented with a sequence of virtual balloons and have to repeatedly decide whether to take a risk by “pumping up” the balloon and potentially burst the balloon or collect the already accumulated sum. Each successful pump increases the amount of reward in the temporary bank but also the probability of a balloon to burst. A trial ends either by a balloon burst, in which case the temporary bank gets emptied and participants lose their earnings from the trial, or if participants choose to collect and transfer their earnings from the temporary to a permanent bank. The goal for participants is to maximise the reward earned by the end of the experiment. A major advantage of the BART lies in its external validity; the adjusted score (mean number of pumps for unexploded balloons) has been repeatedly associated with real life risk-taking behaviours such as smoking or substance use (Aklin et al., 2005; Lejuez et al., 2003; Wallsten et al., 2005).

Until recently, there was no model to reliably estimate differential learning rates in the BART. The Bayesian Sequential Risk-Taking (BSR) model (Pleskac, 2008), originally referred to as “model 3” by Wallsten and colleagues (2005), has been the most prominent model for the BART. Although the original, four-parameter BSR model incorporates learning, its parameters representing initial belief and updating of subjective burst probability were found to be unreliable estimates (Pleskac, 2008; van Ravenzwaaij et al., 2011). This led to a simplification of the model, dubbed BSR-2, with only two parameters indexing the decision maker’s risk-taking propensity and behavioural consistency. Although these two parameters can be recovered reliably, the BSR-2 (and the original BSR model) makes two simplifying assumptions that limit its applicability to a variety of BART paradigms. First, it presumes that burst probabilities are constant across all steps of balloon inflation, suggesting that the model may not generalise well to paradigms with gradually increasing burst probabilities across pumps. Second, participants are assumed to be informed about burst probability, which is incompatible with both real life decision making and most studies where participants are expected to learn through trial and error. As a final drawback, whilst the risk-taking propensity parameter in the BSR model showed good external validity against real-life risk-taking measures, it provided little information about risk-taking propensity beyond that of the adjusted score, which is significantly easier to derive (Wallsten et al., 2005).

To capture differential learning in the BART, Zhou and colleagues (2021) developed the hierarchical Scaled Target Learning (STL) model, which characterises both participants' risk-taking propensity and the extent to which they learn from past experiences. The model encompasses four parameters, which estimate participants' target number of pumps, behavioural consistency, and the degree to which they adjust their target number of pumps in response to positive and negative feedback. The extension of STL, the Scaled Target Learning model with Decay (STL-D), includes an additional decay parameter that estimates how fast participants' adjustments of their target number of pumps decay across trials.

Zhou and colleagues found both their models to have satisfactory parameter recovery and predictive accuracy, with STL-D outperforming STL in most data sets. Furthermore, STL and STL-D's parameter estimate for the target number of pumps and behavioural consistency showed improved external validity compared to the adjusted BART score and the corresponding parameters in the BSR and BSR-2 models, suggesting an improved ability to capture individual differences in risk-taking propensity. Crucially, both learning parameters in the STL(-D) showed good external validity, implying that they adequately characterise learning from one trial to the next. When comparing the STL(-D) against the BSR and BSR-2 models, the former outperformed both models in terms of parameter recovery, predictive accuracy, and external validity. With improved model performance compared to the prominent BSR models, STL(-D) seems a promising tool for improving our understanding of the complex psychological processes underlying the BART, including both risk-taking propensity and differential learning.

Current study

In this study, we fit hierarchical STL and STL-D models to a modified version of the BART paradigm to investigate the degree to which learning bias characterises sequential decision making under different levels of uncertainty. Each participant completed three phases of the BART, characterised by different levels of burst probabilities leading to variable levels of perceived uncertainty across observers. To increase ecological validity, burst probabilities exponentially increased across pumps, mirroring real balloons that are more likely to explode with more and more inflation. First, all participants completed a baseline phase, characterised by an intermediate level of balloon burst probability function, followed by a lucky or an unlucky phase. The lucky phase had a more moderate and the unlucky phase had a steeper increase in their respective balloon burst probability functions compared to the baseline phase. To measure potential order effects that may have confounded behaviour, the order of the lucky and unlucky phases was counterbalanced across participants.

Since human participants flexibly adapt their decision making under different levels of risk and uncertainty in the BART (Kóbor et al., 2023), we did not expect significant differences in participants' risk-taking propensity or learning between the order in which each phase was completed. Given the well-established phenomenon whereby learning rates surge with increasing levels of environmental uncertainty (Behrens et al., 2007; Browning et al., 2015; Palminteri et al., 2017), we expected the learning rates in the unlucky phase to exceed those in the lucky condition. Crucially, evidence for this effect would provide further support that the learning parameters in STL(-D) accurately capture learning in the BART, which was the motivation for developing these models. To the best of our knowledge, our study is the first application of these models to a modified BART paradigm.

In line with results indicating a learning bias in instrumental learning tasks (den Ouden et al., 2013; Frank et al., 2007; Harada, 2020; Lefebvre et al., 2017; Niv et al., 2012; Palminteri, 2023; Palminteri et al., 2017), we expected that a learning bias would also characterise behaviour in the BART. Additionally, we were also interested in how learning bias is related to risk-taking propensity and performance, i.e., the amount of points earned during the experiment. Given that increased learning from positive compared to negative outcomes has been associated with overestimating the value of the riskier decision alternative (Niv et al., 2012), we hypothesised that the magnitude of learning rate bias would be positively associated with risk-taking propensity. However, results regarding the relationship between learning rate bias and performance have been mixed. Whilst Harada (2020) found a negative association between these measures, Lefebvre and colleagues (2017) reported no difference in performance between participants with and without a learning bias. At the same time, Palminteri et al. (2017) reported a negative association between learning bias and performance only following reversals in reward contingencies but not during a stable period of a two-armed bandit task. By investigating how learning bias is linked to performance under different levels of uncertainty, we aim to clarify the conditions under which learning bias affects performance.

Method

Participants

A total of *N*=50 student participants (age: *M* = 21.3 years, *SD* = 2.5 years) took part in the experiment, who were recruited from university courses (for *a priori* power analysis, see the supplementary methods & results section). Participants were randomly assigned to either of the order conditions (Order 1: 6 males, 19 females, Order 2: 4 males, 21 females, Order 1: *M* = 21.6 years, *SD* = 2.8 years, Order 2: *M* = 21 years, *SD* = 2.2 years). All participants had normal or corrected-to-normal vision, reported no existing psychiatric or neurological conditions, and were not taking psychoactive medication at the time of the experiment. Before enrollment in the study, all participants provided written informed consent. The experiment was approved by the United Ethical Review Committee for Research in Psychology (EPKEB) in Hungary, and was conducted in accordance with the Declaration of Helsinki. Participants received course credits in exchange for participation as well as a supermarket voucher. Whilst participants were told that the value of this voucher would vary between 1000-2000 HUF (the equivalent of €2.5 - €5) depending on their performance in the task, all participants received a voucher worth 2000 HUF at the end of the study. All participants were retained in all of our analyses.

Stimuli and task

We utilised a modified version of the Balloon Analogue Risk Task (BART; Fein & Chang, 2008; Kóbor et al., 2015; 2023) to explore whether a learning bias characterises behaviour during sequential decision making under uncertainty. The BART was first proposed by Lejuez and colleges (2002) and has since been established as a widely used and well-validated measure of risk-taking propensity (Aklin et al., 2005; Lejuez et al., 2003). The modified version of the task allowed for shorter trial lengths, which kept the duration of the experiment within a reasonable time range. Furthermore, the implementation of incremental potential reward values allowed for a more accurate assessment of individual differences in risk-taking propensity (Éltető et al., 2019). The task was implemented in Presentation (Version 21.1, Neurobehavioral Systems, Inc., Berkeley, CA) and responses were recorded via a Cedrus RB-540 response device (Cedrus Corporation, San Pedro, CA).

During the task, participants had to repeatedly decide whether to continue ("pump") or stop inflating a virtual balloon that could either increase in size or explode following each inflation step. Successful balloon pumps increased the size of the balloon as well as the reward, but also the likelihood of a balloon burst. Participants used two response keys of the response device to indicate their decision to further pump a balloon or finish the trial and collect their accumulated score from the trial (cash-out). A balloon inflation could result in two outcomes; the balloon would increase in size together with the accumulated score (positive feedback) or the balloon would burst (negative feedback). In case participants decided to stop inflating the balloon, their score earnt in the trial would be transferred to a virtual permanent bank. If a balloon burst ended the trial, the accumulated score in that trial was lost without a decrease in the participant's score in the permanent bank. Participants were instructed to maximise their total score in the task, reflected by the accumulated score in their virtual permanent bank.

During the task, participants could continuously see their accumulated score in the current trial, which was displayed in the middle of the balloon. The accumulated score in the permanent bank, the score collected in the previous trial, and the response key options for inflating the balloon and collecting the accumulated score were also displayed throughout the experiment. The feedback of a balloon burst was represented by a fragmented balloon and the cash-out screen informed participants about the score they earnt in the trial (Fig.1). Each feedback screen was presented for 3000 ms and participants' responses were not limited in time.

Participants completed a total of 270 trials in the experiment, divided into three 90-trial phases, each characterised by different balloon burst probabilities. The baseline phase was characterised by an intermediate level of balloon burst probability, which probability increased in the unlucky phase and decreased in the lucky phase. Each participant started the task with a baseline phase, after which half of the participants first completed the lucky phase followed by the unlucky phase (Order 1) or continued with the unlucky phase and finished the task with the lucky phase (Order 2). In the baseline and lucky phases, the maximum number of balloon inflation steps was 20, whilst this was limited to 10 in the unlucky phase.

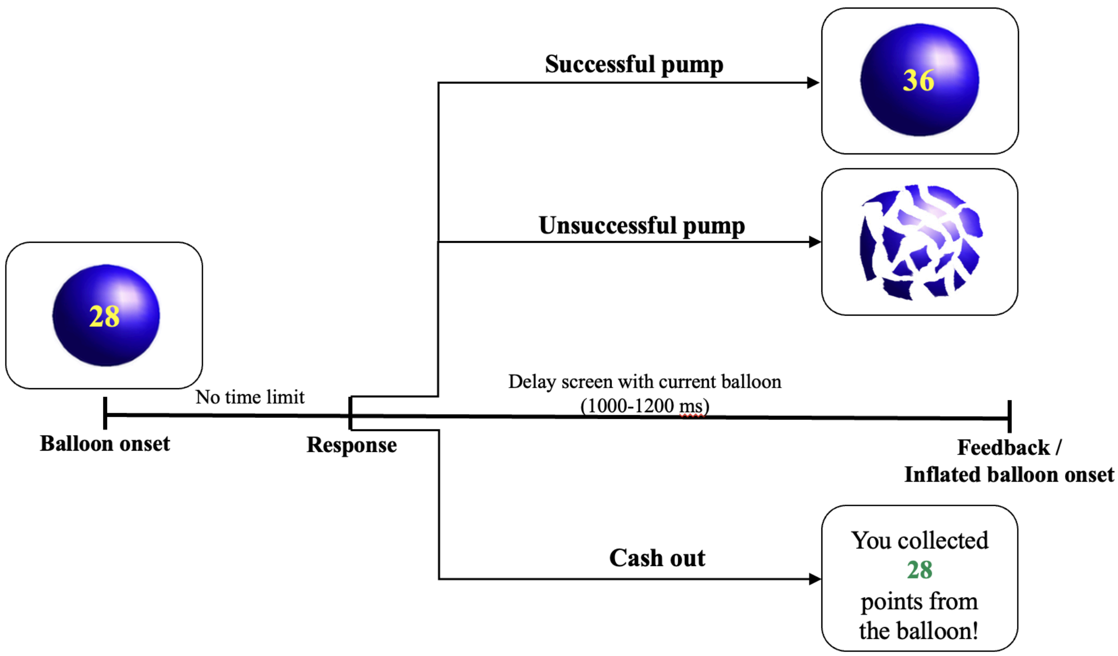


Figure 1. An example trial of the modified BART. Participants had to repeatedly decide whether to continue inflating a virtual balloon, which could either increase in size or explode, or stop inflating and cash out their accumulated score from the trial. Each inflation step increased the balloon's size and the reward as well as the likelihood of a balloon burst (Eq.1). Participants could see their score accumulated in the current trial in the balloon and had no time limit to respond.

The balloon would not explode following the first two inflations in any of the phases, after which point the probability of balloon burst increased. The burst probabilities for each inflation steps were determined according to

$$P(e_k) = 0, \quad \textit{if } p_k \leq 2, \textit{ in all phases}$$

$$P(e_k) = \frac{1}{21 - p_i}, \quad \textit{if } 3 \leq p_k \leq 19 \textit{ in the baseline phase}$$

$$P(e_k) = \frac{1}{31 - p_i}, \quad \textit{if } 3 \leq p_k \leq 19, \textit{ in the lucky phase}$$

$$P(e_k) = \frac{1}{11 - p_i}, \quad \textit{if } 3 \leq p_k \leq 9 \textit{ in the unlucky phase}$$

$$P(e_k) = 1, \quad \textit{if } p_k = 20 \textit{ in the baseline and lucky phases}$$

$$P(e_k) = 1, \quad \textit{if } p_k = 10 \textit{ in the unlucky phase,} \qquad (1)$$

where $P(e_k)$ is the probability that the balloon explodes on the *k*th inflation step, and $npump_k$ represents the pump (i.e., inflation) number within trial *k*. Thus, the balloon would explode on the 20th pump in the baseline and lucky phases and on the 10th pump in the unlucky phase (Fig.2). As a result of the distinct burst probabilities in the different phases of the experiment, the optimal number of pumps differed in each phase. For baseline balloons, it was most advantageous to inflate the balloon 13 times, whilst the highest expected return was associated with 19 and 6 pumps in the lucky and unlucky phases, respectively (for more details, see the Supplementary Methods in Kóbor et al., 2023). Participants were naïve regarding the burst probabilities in the experiment, including the zero probability of balloon burst in the first two inflation opportunities. Participants were also unaware that burst probabilities would change during the experiment, and the beginning of a new phase was not signalled to participants.

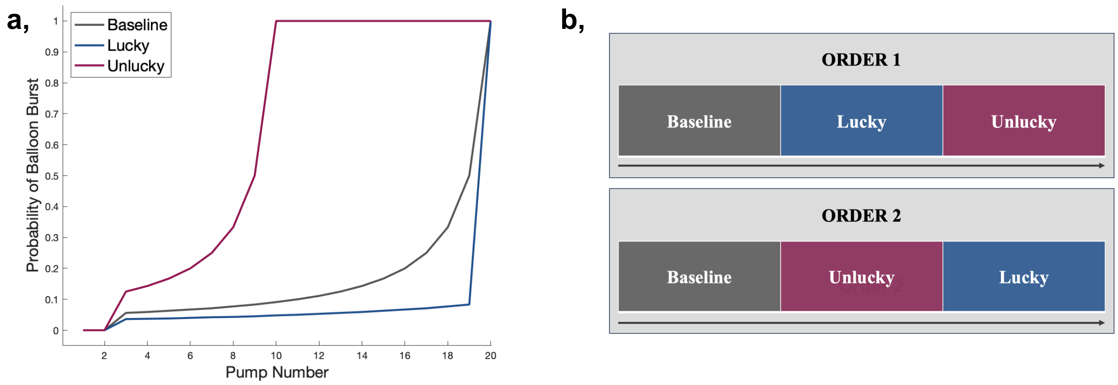


Figure 2. Experimental design a, Illustration of the balloon burst probabilities in each phase of the experiment. Balloon bursts were disabled for the first two balloon pumps in each phase, after which burst probability was controlled by a separate truncated power function for each phase (Eq.1). The balloon was certain to explode on the 20th pump in the baseline and lucky phases and on the 10th pump in the unlucky phase. **b,** Representation of the two experimental groups. In both groups, participants first completed 90 trials of baseline balloons. After these, participants in Order 1 faced 90 lucky balloons, followed by 90 unlucky balloons. Order 2 counterbalanced the order of the lucky and unlucky balloons; baseline balloons were succeeded by a set of 90 unlucky balloons, then a set of 90 lucky balloons.

For each successful balloon inflation within trial *k*, the number of points participants earned equaled the latest pump number $npump_k$. Thus, participants could gain one point for the first successful balloon inflation, two for the second successful inflation (with an accumulated score of 3 in the trial), three for the third successful inflation (with an accumulated score of 6 in the trial), and so on. Participants' overall score was calculated as the sum of points earned in each trial of the experiment.

Participants could freely decide whether to inflate the balloon or collect their accumulated score in 80 trials in each phase. In the remaining 10 trials of each phase, participants were instructed to either inflate the balloon to a predetermined point or until it exploded. These forced-choice trials were included in the experiment to guide participants towards the optimal number of pumps in each phase, and were presented in the same predetermined trial positions to all participants in order to control for across-participant variance.

Procedure

During recruitment, participants took part in a neuropsychological assessment in order to evaluate potential factors that could alter their performance in the BART, including impulsivity, depression, or anxiety. Selected participants took part in two separate experimental sessions. On the first day, participants were asked about any existing medical conditions and medication regimens, their consumption of cognitive performance enhancing drugs, and their motor skills and alertness levels were evaluated. Participants' emotional affect and cognitive performance such as executive functions and working memory were also evaluated. These measurements were collected to pursue hypotheses not explored in this work.

On the second day, participants were questioned about factors that could affect their cognitive performance such as sleep quality, mood, alertness, and whether they consumed cognitive performance enhancing drugs. Additionally, participants' emotional affect was also evaluated. Before beginning the BART, participants had the chance to practise the task with the experimenter, which included six forced-choice balloon trials. During the main task, participants could take predetermined short breaks every 20-25 trials, and there was an additional larger break halfway through the experiment. The BART was followed by a short verbal interview to assess participants' strategies throughout the task and the degree to which they had awareness of the presence of the different phases in the experiment. These results are not described in this study. During the task, continuous electroencephalogram (EEG) data were also recorded. As the analysis of electrophysiological data is outside the scope of this study, details on the recording and analysis of the EEG data are omitted. Altogether, the second experimental session took approximately 2-2.5 hours.

Computational modelling

The Scaled Target Learning (STL) model (Zhou et al., 2021) characterises learning in the BART through adjustments in participants' number of pumps; positive feedback increases, whilst negative feedback decreases the number of pumps. This kind of learning originates from The Law of Effect (Thorndike, 1898), according to which people are prone to repeat choices that have resulted in desirable outcomes and tend to scale down choices that have led to undesirable outcomes. Therefore, STL predicts that participants would increase their target number of pumps following the collection of a reward, and reduce their target number of pumps following an unwanted balloon burst. Crucially, STL implements separate learning rates for wins (*vwin*) and losses (*vloss*) to account for the distinct degrees of sensitivity to rewards and punishments (Cazé & van der Meer, 2013; Corr, 2004; Frank et al., 2007; Gray, 1975; Lefebvre et al., 2017; Niv et al., 2011; Sharot et al., 2011) and the differential neural mechanisms that implement approach and avoidance learning (Daw et al., 2002; Fouragnan et al., 2015; O'Doherty et al., 2001; Schultz, 2016; Palminteri & Pessiglione, 2017; Seymour et al., 2007).

STL does not assume that an intrinsic risk-taking propensity guides behaviour in the BART. Instead, it assumes that participants begin the task with a target number of pumps (ω_k) in mind and adapt this value after each trial according to

$$\omega_k = \omega_{k-1} \times \left(1 + vwin \cdot \frac{npump_{k-1}}{nmax}\right), \qquad \textit{if participant collects in trial k-1}$$

$$\omega_k = \omega_{k-1} \times \left(1 - vloss \cdot \left(1 - \frac{npump_{k-1}}{nmax}\right)\right), \quad \textit{if balloon explodes in trial k-1} \qquad (2)$$

with *vwin*, *vloss* > 0. In STL, ω_k is scaled by the design parameter *nmax*, representing the maximum pump number possible in each trial, so that the value of ω_k falls between 0 and 1. This was implemented in order to account for two phenomena observed in the BART. First, participants tend to increase their pumps following a win with a larger compared to a smaller reward value. Second, participants tend to pump more following a loss with a higher compared to a lower reward that could have been obtained (Schmitz et al., 2016; Zhou et al., 2021). Thus, adjustments after a win ($vwin \cdot \frac{npump_k}{nmax}$) imply a larger increase in ω_k after a larger collection in the previous trial ($\frac{npump_{k-1}}{nmax}$), whilst adjustments after a loss ($vloss \cdot (1 - \frac{npump_k}{nmax})$) imply a smaller reduction in ω_k following a loss with a larger potential reward ($\frac{npump_{k-1}}{nmax}$). Additionally, as the amounts of reward collected or lost due to a balloon burst are scaled by *nmax*, model estimates across various experimental designs of the BART (i.e., different burst probabilities) can be directly compared (Zhou et al., 2021).

STL further assumes that human behaviour entails a degree of randomness; participants' decisions are probabilistic and are not solely based on their target number of pumps ω_k, but are also determined by participants' behavioural consistency β which influences the degree to which participants behave rationally. Thus, the probability that participants will pump on trial *k* for a given pump opportunity *l (= 1, 2, …)* is given by

$$P_{kl}^{pump} = \frac{1}{1+e^{\beta \cdot (l-\omega_k)}} \, , \tag{3}$$

with $\beta \geq 0$. Thus, *l* increases with more pumps on trial *k*, and the probability that participants will further pump declines until *l* reaches the target number of pumps ω_k, when the probability of pumping equals chance. Since participants with higher β rely more on their target number of pumps ω_k, behavioural consistency β can be understood as participants' prior evaluation of options (Wallsten et al., 2005).

The Scaled Target Learning with Decay (STL-D) model builds on STL by including an additional decay parameter α, which reflects how fast adjustments in ω_k decay across trials. STL-D characterises decay as a linear function according to

$$\omega_k = \omega_{k-1} \times \left(1 + \frac{vwin \cdot \frac{npump_{k-1}}{nmax}}{1+\alpha \times (k-1)}\right), \qquad \textit{if participant collects in trial k-1}$$

$$\omega_k = \omega_{k-1} \times \left(1 - \frac{vloss \times (1 - \frac{npump_{k-1}}{nmax})}{1+\alpha \times (k-1)}\right), \textit{ if balloon explodes in trial k-1} \qquad (4)$$

with *vwin*, *vloss*, α > 0. Similarly to STL, STL-D assumes that participants adjust their target number of pumps as a function of past outcomes. However, the degree of adjustment decreases across trials *k* and is reflected by the decay parameter α. Finally, both STL and STL-D assume the same choice process, given by Eq.3, whereby ω_k controls the probability of pumping P_{kl}^{pump} given each pumping opportunity *l*. Thus, whilst STL has four free parameters reflecting participants' target number of pumps ω_k, behavioural consistency β, and learning rates following wins and losses (*vwin* and *vloss*, respectively), STL-D has a fifth free parameter α reflecting decay.

Model fitting, comparison, and parameter validity checks

We implemented hierarchical Bayesian analysis (Gelman et al., 2013; Zhou et al., 2021) to estimate individual and group-level parameters for the STL and STL-D models, whereby individual-level parameters were drawn from normally-distributed group-level distributions with weakly informative priors. Analysis was performed in the *Rstan* package (version 2.17.2; Stan Development Team, 2019) in R (version 3.3.3; R Core Team, 2019) and utilised a Hamiltonian No U-Turn sampler (NUTS) to derive the joint posterior distribution of parameters. For each model, we generated 5000 samples after discarding the first 1000 observations as burn-in.

We fit both the STL and STL-D model to each of the three phases within each order, resulting in 6 separate fits in total (Order 1 baseline, Order 2 baseline, Order 1 lucky, Order 2 lucky, Order 1 unlucky, Order 2 unlucky) for each model. We used the 80 free-choice trials within each phase as we considered the inclusion of the forced-choice trials in the model conceptually problematic as participants had to carry out external instructions in these trials. Nevertheless, when the models were implemented utilising all 90 trials of each phase, changes in the resulting parameter estimates and model fits were negligible. We monitored model convergence by calculating the $\widehat{R}$ statistic (Gelman & Rubin, 1992) to compare within- and between-chain variance across four chains of each model. All model parameters successfully converged with $\widehat{R}$ <1.01 at the group level. Estimated levels of the decay parameter α fell within the recommended range between 0 and 0.1 (Zhou et al., 2021) for the STL-D model. Values of α above this range have been associated with reduced recovery of the learning parameters *vwin* and *vloss*.

We evaluated the predictive accuracy of our models by comparing the leave-one-out information criterion (*LOOIC*; Vehtari et al., 2017). This measure of leave-one-out cross-validation estimates the out-of-sample predictive accuracy of Bayesian models, with lower *LOOIC* values representing improved predictive accuracy. It is considered more accurate compared to other information criteria such as the *Akaike Information Criterion* (*AIC*; Akaike, 1978) or the *Deviance Information Criterion* (*DIC*; Spiegehalter et al., 2002). We computed *LOOIC* via the *loo* R package (Vehtari et al., 2017). This comparison revealed that STL-D fit our data slightly better (Table 1), with lower *LOOIC* values for 4 out of the 6 phases. Consequently, we used the parameter estimates from STL-D for further analyses.

Table 1. Model comparison. Leave-one-out information criterion (LOOIC) for the STL and STL-D models in the different experimental conditions. Lower LOOIC values illustrate a better model fit. The last column shows the difference between the LOOIC values associated with each model, i.e., ΔLOOIC = LOOIC(STL) - LOOIC(STL-D). Thus, positive ΔLOOIC values indicate an improved model fit for STL-D compared to STL, whereas negative values illustrate a better model fit for STL compared to STL-D.

Order	Phase	STL	STL-D	ΔLOOIC
Order 1	Baseline	5600	5621	-21
	Lucky	6776	6753	23
	Unlucky	4561	4530	31
Order 2	Baseline	5447	5423	24
	Lucky	6936	6944	-8
	Unlucky	4303	4240	63

Secondary analyses

To analyse the effect of phase and order manipulations, we utilised *lm()* in R (version 3.3.3; R Core Team, 2019) to perform a two-way analysis of variance (ANOVA) on each of the individual-level STL-D parameter estimates. Although directly comparing group-level parameter estimates in a Bayesian framework (see above) would be preferred, this would considerably increase the complexity of our computations. Consequently, we carried out secondary analyses on the individual-level parameters to make the comparisons and their interpretation more straightforward. We baseline-corrected each parameter by subtracting the baseline value from the corresponding parameter estimates from the lucky and unlucky conditions. We utilised Bonferroni-correction to reduce the likelihood of Type I errors in our statistical testing. Consequently, we divided the original *p* value of .05 by the number of tests carried out (5, one for each STL-D parameter) and evaluated effects against an adjusted *p*-value of .01.

As we found no learning rate difference across corresponding phases across the two order conditions, we combined parameter estimates phase-wise in all subsequent analyses. To examine whether a learning rate bias is present in our data, we quantified learning rate bias by normalising the learning rate difference (Niv et al., 2012; Palminteri et al., 2017) for each participant and phase according to

$$bias = \frac{vwin - vloss}{vwin + vloss}. \tag{5}$$

Finally, we examined how individual-level learning bias was linked to risk-taking propensity and performance. We used the adjusted score (mean number of pumps across unexploded balloons) and the total number of points earned by each participant as a proxy for risk-taking propensity and performance, respectively. We used across-participant Pearson's correlations to evaluate the degree of association across these variables separately for each experimental phase. In line with the findings by Niv and colleagues (2012), we hypothesised that there would be a positive link between learning bias and risk-taking propensity (Fig.6b). Due to the mixed results regarding the association between learning rate bias and performance (Lefebvre et al., 2017; Palminteri et al., 2017; Harada, 2020), these significance tests were undirected (Fig.6c).

Results

Adjusted Score and Performance

To evaluate behaviour across the different experimental phases, we calculated each participant's adjusted score (i.e., the mean number of balloon inflations on unexploded balloons) and the total points earned. There were no significant differences in either of these measures between corresponding phases across the two orders ($p > .05$ for all two-tailed t-tests). Consequently, we aggregated data across the two orders to examine behavioural differences across conditions.

As Figure 3a shows, the adjusted score was significantly higher in the lucky compared to both the baseline ($t(98) = -2.79$, $p = .006$, 95% $CI = [0.37, 2.20]$) and unlucky ($t(98) = 14.43$, $p < .001$, 95% $CI = [4.90, 6.47]$) conditions. As expected, the adjusted score was significantly higher in the baseline compared to the unlucky condition ($t(98) = 15.14$, $p < .001$, 95% $CI = [3.82, 4.98]$). Figure 3b illustrates that participants achieved a significantly higher number of points in the lucky than in the baseline ($t(98) = -7.25$, $p < .001$, 95% $CI = [3.82, 4.98]$) or unlucky ($t(98) = 23.22$, $p < .001$, 95% $CI = [897.13, 1573.47]$) conditions, with significantly more points earned in the baseline compared to the unlucky phase ($t(98) = 17.51$ $p < .001$, 95% $CI = [1403.47, 1665.73]$).

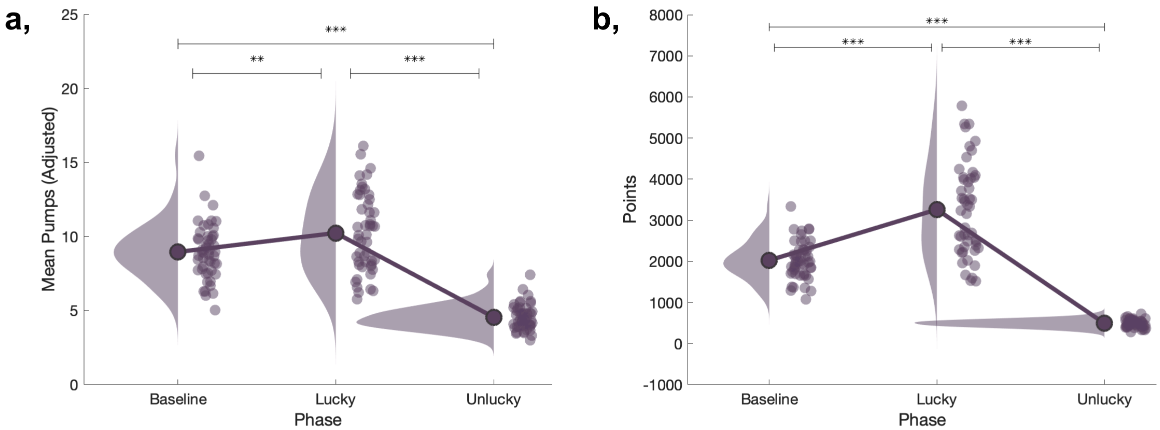


Figure 3. Adjusted score and earnings. a, Adjusted score in each condition, aggregated across the two experimental orders. The adjusted score was calculated for each participant separately for trials in which the balloon did not burst. **b,** Points earnt by each participant across the different experimental conditions. We aggregated data by phase across the two experimental orders.

Modelling results

We implemented Hierarchical Bayesian Analyses (Gelman et al., 2013; Zhou et al., 2021) to estimate individual and group-level parameters of the STL and STL-D model separately for each experimental phase of each order group, resulting in six independent runs per model. The two models produced similar parameter estimates and model fits for each run. However, as Table 1 shows, STL-D slightly outperformed STL, with lower leave-one-out information criterion (LOOIC; Vehtari et al., 2017; Zhou et al., 2021) values for 4 out of the 6 phases. Consequently, we utilised the parameter estimates from STL-D for further analyses. Posterior distributions of all group-level STL-D parameters, broken down by phase and order type, are displayed in Figure 4.

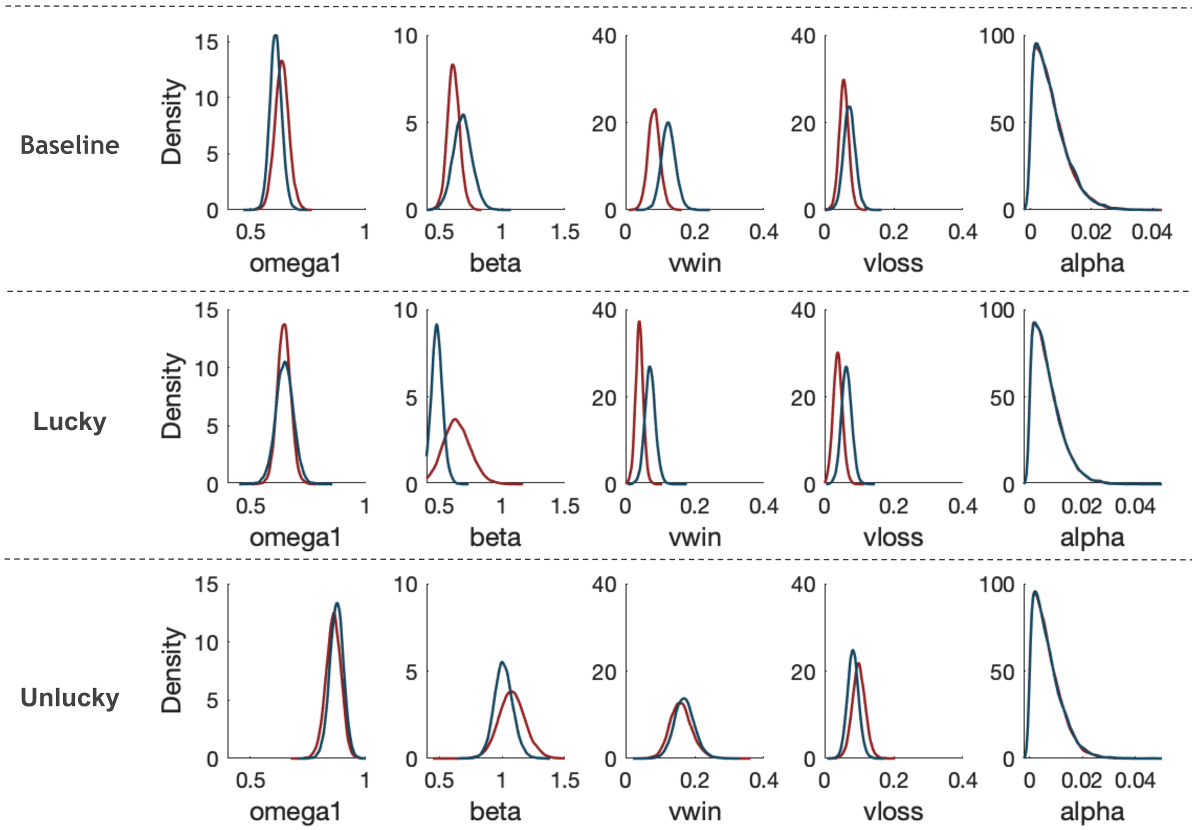


Figure 4. Posterior distributions of STL-D parameter estimates. Group-level posterior distributions are shown in three rows for each phase and in five columns for each parameter. The red and blue lines indicate posterior distributions for Order 1 and Order 2, respectively.

Phase and order effects

To evaluate potential differences across the experimental phases and orders, we performed a 2 x 2 mixed ANOVA on each of the mean individual parameter estimates from the STL-D model. We baseline-corrected parameter estimates from the lucky and unlucky phases by subtracting their corresponding baseline value. The baseline-corrected, individual parameter estimates, broken down by phase and order, are depicted in Fig.5. Each effect was evaluated against the Bonferroni-corrected *p*-value of .01. For the parameter estimating participants' target level of pumps ω_k, we found a main effect of phase ($F(1,48) = 75.75$, $p < .001$). We did not find a main effect of order ($F(1,48) = 1.81$, $p = .18$) or an interaction effect ($F(1,48) = .05$, $p = .82$). For the parameter β, reflecting participants' behavioural consistency, we identified a main effect of both phase ($F(1,48) = 55.62$, $p < .001$) and order ($F(1,48) = 8.67$, $p = .004$), without a significant interaction ($F(1,48) = .48$, $p = .49$). Please note that both ω_k and β are scaled by the maximum possible number of pumps *nmax* for each phase, which differed in the lucky and unlucky phases. Whilst comparison within the STL(-D) model is possible across conditions with different maximum burst points, large *nmax* differences may distort results.

We found a similar pattern for how learning from wins and losses, captured by the parameters *vwin* and *vloss*, respectively, changed throughout the task. Specifically, there was a main effect of phase for both *vwin* ($F(1,48) = 44.21$, $p < .001$) and *vloss* ($F(1,48) = 21.32$, $p < .001$), without a significant main effect of order for either *vwin* ($F(1,48) = 1.76$, $p = .19$) or *vloss* ($F(1,48) = 1.85$, $p = .18$). The interaction effect was not significant for either *vloss* ($F(1,48) = 5.83$, $p = .018$) or *vwin* ($F(1,48) = .45$, $p = .51$) at an adjusted significance level of $p = .01$. Similarly to the learning parameters, an equivalent ANOVA on the decay parameter α revealed a significant main effect of phase ($F(1,48) = 14.76$, $p < .001$), without a significant effect of order ($F(1,48) = .84$, $p = .84$) or a significant interaction effect ($F(1,48) = .56$, $p = .46$). Larger values for the learning parameters in the unlucky compared to lucky phase provide evidence for the external validity of these parameters as higher levels of environmental volatility, as introduced by the unlucky phase, have been associated with increased learning rates (Behrens et al., 2007; Browning et al., 2015).

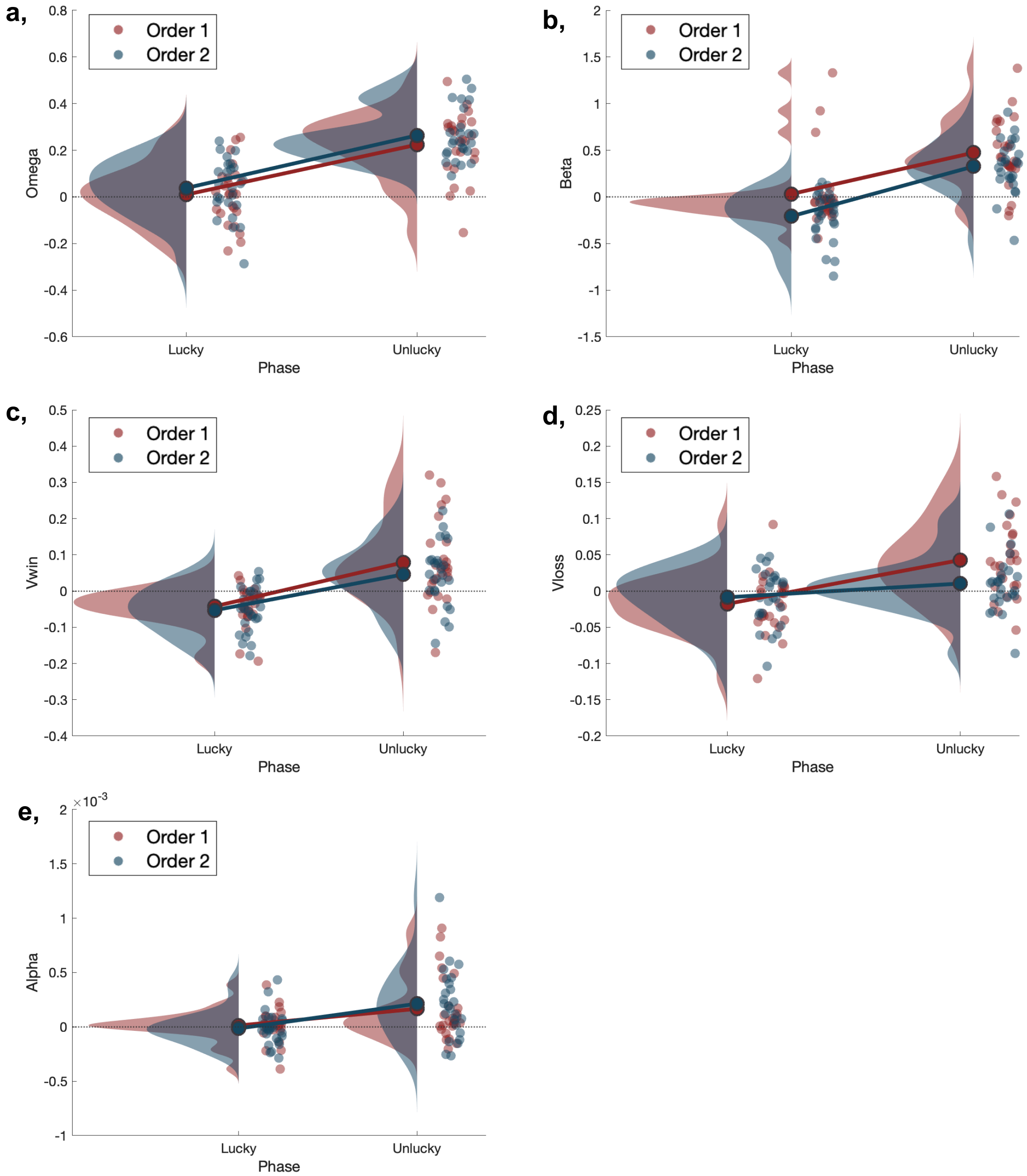


Figure 5. Phase and order effects. Individual parameter estimates for the target number of pumps ω_k (**a**), behavioural consistency β (**b**), learning from wins *vwin* (**c**) and losses *vloss* (**d**), and decay (**e**) from the Scaled Target Learning Model with Decay (STL-D) are shown separately for each phase (lucky and unlucky) and order (Order 1, Order 2). All parameters were baseline-corrected by subtracting the parameter estimates linked to the baseline phase from the corresponding parameter estimates from the lucky and unlucky conditions.

Learning bias

We utilised both individual- and group-level (see Supplementary section) analyses to determine whether learning bias is present in the BART under varying levels of uncertainty. First, we carried out one-tailed, undirected t-tests assessing whether the participant-wise normalised learning rate difference (Eq.5) is significantly different from zero in each phase of the experiment (Fig.6a). When evaluated against an adjusted *p*-value of .017, we found that the normalised learning bias was significantly higher than zero in the baseline (M = .25, SD = .21, $t(49)$ = 8.30, p < .001, 95% CI = [.19, .31]) and unlucky (M = .28, SD = .31, $t(49)$ = 6.40 , p < .001, 95% CI = [.19, .36]) phases, but not in the lucky phase (M = .07, SD = .24, $t(49)$ = 2.07, p = .04, 95% CI = [.002, .14]).

To examine the link between learning bias and risk-taking propensity as well as performance, we established Pearson correlations across participants. To measure risk-taking propensity, we utilised the adjusted score, i.e., the mean number of pumps in trials with unexploded balloons. We quantified participants' performance as the number of points earned in each phase. We correlated the participant-specific adjusted score and the number of points earned with individual-level estimates of the normalised learning rate difference (learning rate bias, Eq.5) in each phase.

As shown in Fig.6b, we found that the magnitude of the learning bias was positively associated with the adjusted score in all phases (Fig.6b; baseline: $r(48)$ = .53 , p < .001, lucky: $r(48)$ = .35 , p = .01, unlucky: $r(48)$ = .47, p < .001). In line with previous results (Niv et al., 2012), this suggests that participants with larger learning bias can be characterised by increased risk-taking propensity. Additionally, performance significantly and negatively correlated with learning bias in the unlucky phase ($r(48)$ = -.51, p < .001), whilst this association was not significant in the baseline ($r(48)$ = .21, p = .14) and lucky ($r(48)$ = .11, p = .43) phases (Fig.6c).

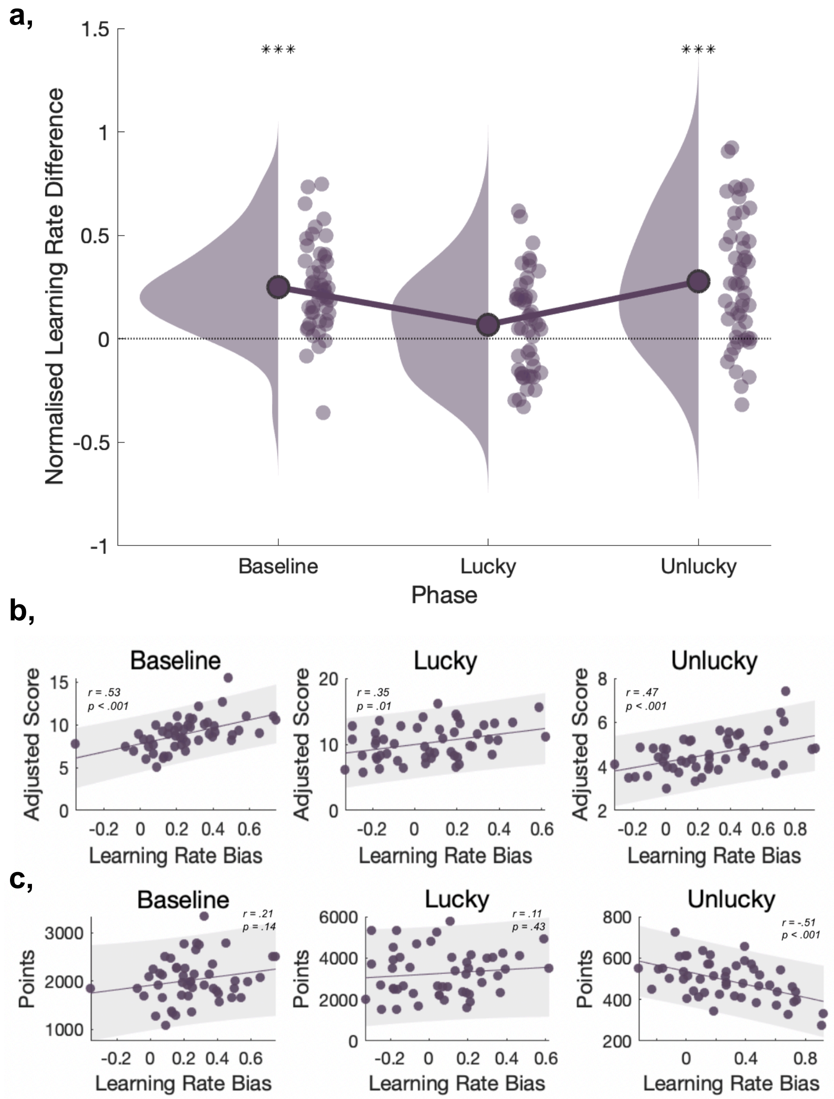


Figure 6. Learning bias, risk-taking propensity, and performance. a, Participant-wise normalised learning rate difference is shown for each phase. One-way t-tests revealed a significant bias in the baseline and unlucky phases, but not in the lucky phase. Significance was evaluated against the Bonferroni-corrected *p*-value, adjusted by the number of tests carried out. Across-participant Pearson's correlation between individual-level normalised learning bias (Eq.5) and risk-taking propensity (**b**) as well as performance (**c**) in each phase of the experiment. Risk-taking propensity was quantified as the adjusted score, i.e. mean number of pumps on unexploded balloons. Performance was measured by the total number of points earned in each phase. The Pearson's correlation coefficient and its corresponding *p*-value are shown on the top right of each graph. Data were aggregated across corresponding phases of Orders 1 and 2.

Discussion

We provide evidence for learning bias in sequential decision making under uncertainty, which appears to be dependent on increased levels of perceived uncertainty. Moreover, learning bias was negatively associated with performance only under the highest level of uncertainty, further implying the important modulatory role of uncertainty. Additionally, learning bias and risk-taking propensity were positively associated in all conditions, implying that the degree of learning bias may universally shape risk-taking preferences. To our knowledge, this study constitutes the first application of the STL and STL-D models since their original development (Zhou et al., 2021), with both models appearing to accurately capture both risk-taking propensity and the learning process in a modified version of the BART.

Learning bias

The presence of learning rate bias under increased uncertainty is consistent with previous studies reporting learning bias during instrumental learning (den Ouden et al., 2013; Frank et al., 2007; Lefebvre et al., 2017; Niv et al., 2012; Palminteri, 2023; Palminteri et al., 2017). Whilst Harada (2020) found learning bias in the Iowa Gambling Task, a similar paradigm to the BART compared to two-armed bandit tasks, when estimating static learning rates, this bias disappeared with the introduction of time-varying learning rates in their Q-learning model. Although STL-D does not estimate learning rates for each trial, it models learning as a non-stationary, decaying process with linearly decreasing learning rates across trials. Consequently, our results provide evidence that learning bias is not merely a by-product of static learning rates. Overall, results from both the current and previous (den Ouden et al., 2013; Frank et al., 2007; Harada, 2020; Lefebvre et al., 2017; Niv et al., 2012; Palminteri, 2023; Palminteri et al., 2017) studies indicate that learning bias is a universal phenomenon in human reward learning.

Both our Bayesian analyses at the group-level (Supplementary Fig.1) and our frequentist analyses at the individual-level (Fig.6) implicated learning bias in the unlucky but not in the lucky condition. The two lines of analyses diverged when it comes to the baseline phase; the individual-level analysis suggested that learning bias was present in this phase, whilst the group-level analysis could not credibly confirm this. It is worth noting that group-level comparison is inherently more conservative as it reflects the behaviour of an entire group, including participants with both low and high bias. At the same time, although drawn from a population distribution, participant-specific parameters can accommodate individual differences in behaviour. As such, individual-level analyses may be more accurate in capturing participants' underlying behaviour, implying that learning bias was indeed present in the baseline phase.

Despite learning bias appearing as a robust phenomenon across different instrumental learning tasks, its purpose and behavioural implications have remained elusive. Palminteri et al. (2017) and Palminteri's (2023) meta-analysis showed that learning bias in two-armed bandit tasks arises from a confirmation bias, rather than a positivity bias. In other words, participants seemed to preferentially learn from positive outcomes because the outcome confirms their choice strategy, not on the grounds that they are positively valenced. However, these studies utilised cognitive models with static learning rates, which may not be perfectly suited to account for the stationary reward contingencies of the task. As current BART models do not allow for the differentiation of confirmation or valence-induced bias, meaningful assessment awaits the development of further paradigms and cognitive models.

In line with the account that reinforcement learning is inherently related to risk-sensitivity (Niv et al., 2012), our results indicate a consistent positive association between the adjusted score, used to index risk-taking propensity (Aklin et al., 2005; Lejuez et al., 2003; Wallsten et al., 2005), and learning bias across all experimental conditions (Fig.6b). At the same time, both learning rates showed a negative relationship with the adjusted score (Supplementary Fig.2). These results are consistent with previous findings (Niv et al., 2002; 2012) and can be explained within reinforcement learning theory. Specifically, higher learning rates cause more fluctuation in estimated value and therefore lead to more risk-aversion. If positive feedback increases stimulus value more than negative feedback decreases it, then stimulus value will be higher than the mean nominal outcome, leading to increased risk-seeking. Consequently, biassed learning towards positive outcomes leads to an overestimation of values, resulting in a higher target number of pumps, as borne out by our results. Our results also indicated that compared to *vwin*, *vloss* is more consistently and more strongly linked to risk-taking propensity. If learning bias indeed arises from reduced negative feedback processing (Lefebvre et al., 2017), it is plausible that the inverse relationship between *vloss* and risk-aversion on the one hand, and learning bias and risk-aversion on the other hand, represent the same cognitive process. Thus, although our study provides evidence that learning bias may underlie risky behaviour, future research should confirm whether this association exists beyond the increased risk-aversion resulting from reduced negative outcome processing.

Our results also confirm previously reported associations between reward learning and performance. First, we found that increased learning from positive and negative feedback (except in the unlucky phase) was associated with lower performance across the different experimental phases (Supplementary Fig.3). This is consistent with the notion that a slower integration of outcomes is necessary for the generalisation of probabilistic reward values (Frank et al., 2007). We speculate that increased learning from negative feedback did not remain maladaptive in the unlucky phase as the change in reward contingencies in this condition was indicated by balloon bursts, requiring adaptation that is primarily based on negative feedback.

Additionally, we found that learning bias was only significantly related to performance in the unlucky phase (Fig.6c), where increased bias was associated with reduced performance. This reflects the contradictory results across previous studies whereby learning bias was found to be maladaptive by Harada (2020) but not by Lefebvre and colleagues (2017). Our results exhibit a strikingly similar pattern to those by Palminteri and colleagues (2017) in that higher learning bias was only found to be maladaptive with the introduction of increased environmental uncertainty (reversals). Palminteri and colleagues also showed that this decreased ability to flexibly adapt to changing, uncertain environments results from confirmation bias, whereby participants showed increased perseveration despite obtaining new information from negative feedback. Simulations by Caze and van der Meer (2013) also suggested that the (mal)adaptiveness of learning bias (either in favour of positive or negative feedback processing) depends on environmental attributes such as the rate of reward. Considering these results, it is likely that both the presence and maladaptiveness of learning bias is contingent on environmental features (i.e., uncertainty), perhaps as a means to flexibly adjust one's exploration-exploitation strategy (Harada, 2020). Nevertheless, even if undue optimism has a net negative impact on learning, it may still be a "self-serving" feature of human cognition that promotes self-esteem and confidence, both of which are related to positive life outcomes (Carver et al., 2010; Weinstein et al., 1980).

Computational Modelling of the BART

Our study was made possible due to the recent development of STL and STL-D models by Zhou and colleagues (2021). We successfully applied both models to a modified, BART paradigm with varying levels of burst probability functions (i.e., uncertainty) across conditions. These models were originally developed to reliably and meaningfully characterise learning during sequential decisions. Crucially, the model distinguishes learning from positive and negative feedback by estimating differential learning rates to account for the distinct sensitivity to rewards and punishments (Cazé & van der Meer, 2013; Corr, 2004; Frank et al., 2007; Gray, 1975; Lefebvre et al., 2017; Niv et al., 2011; Sharot et al., 2011) and the separate neural processes facilitating approach and avoidance learning (Daw et al., 2002; Fouragnan et al., 2015; O'Doherty et al., 2001; Schultz, 2016; Palminteri & Pessiglione, 2017; Seymour et al., 2007).

Indeed, the differential learning rates estimated by STL-D reflected the change in response to manipulations of environmental uncertainty in accordance with the well-established phenomenon that learning rates increase under heightened levels of uncertainty (Behrens et al., 2007; Browning et al., 2015; Palminteri et al., 2017). Both Zhou and colleagues' (2021) and our results imply that STL(-D) reliably and meaningfully characterises learning in the BART. This is a major improvement compared to the prominent BSR (Wallsten et al., 2005) and BSR-2 (Pleskac, 2008; van Ravenzwaaij et al., 2011) models as the former cannot reliably recover its learning parameter (Pleskac, 2008; van Ravenzwaaij et al., 2011) and the latter does not take learning into account. Unlike the BSR models, STL(-D) can be applied to paradigms with gradually increasing burst probabilities and does not require participants to be aware of the underlying burst probabilities, and can be consequently applied to a greater variety of experimental paradigms.

Consistent with Zhou and colleagues' (2021) findings, the STL model appears to be improved on by its extension, STL-D, implying that in contrast to assuming constant learning, a linearly decaying learning process better characterises behaviour in the BART. These results parallel other findings in the reinforcement learning literature that indicate improved model fit for Q-learning models including decay (Geana et al., 2022; Radulescu et al., 2016; Yechiam & Busemeyer, 2005). The decay parameter in STL-D indicates how fast adjustments in pumping behaviour decline with experience (Eq.4). That is, higher decay reflects a larger weight of past experiences as participants change their pumping behaviour the most in the beginning of the experiment and adjust their behaviour progressively less across trials. Indeed, given that reward contingencies changed with the introduction of a new experimental phase and each phase is modelled separately, it is adaptive to integrate feedback information over a longer period of time to avoid behaviour being overly influenced by the most recent outcomes throughout the experiment (Frank et al., 2007). Accordingly, the higher decay in the unlucky compared to the lucky phase suggests increased emphasis on learning in the beginning of the phase. Considering that higher decay is adaptive with the introduction of a larger change in reward contingencies (i.e., in the unlucky phase), the decay parameter in STL-D appears to constitute a meaningful addition to modelling the learning process in the BART.

In line with our initial expectation that humans flexibly adapt their decision making in response to the level of environmental uncertainty (Kóbor et al., 2023), we did not find significant differences in the STL-D parameters reflecting participants' target level of pumps, learning rates, or decay across the two orders. Whilst we observed a significant order effect in the behavioural consistency parameter, this was largely driven by three outliers in the lucky phase of Order 1 (Fig.6b), which questions the generality of this effect. Furthermore, our results suggest increased behavioural consistency and target level of pumps in the unlucky compared to the lucky phase. Although this may seem counterintuitive, both parameters are proportional to the maximum number of pumps, which differed across the two conditions. Despite participants pumped less in the unlucky compared to the lucky phase, the lower number of possible pumps in the unlucky phase generated higher parameter estimates for the target number of pumps and behavioural consistency.

This counter-intuitive conclusion likely stems from modifications to the original BART paradigm (Lejuez et al., 2002; Wallsten et al., 2005), which had a substantially higher maximum burst point as well as constant burst probabilities across balloons. In fact, STL and STL-D are applicable to paradigms with gradually increasing burst probabilities and meaningful comparison across conditions with different maximum burst probabilities (Zhou et al., 2021). However, it appears that simultaneous adjustments in these aspects of the task, including large differences in maximum burst points across conditions, may result in a biassed comparison across conditions or experiments. To reliably compare participants' behavioural consistency and target number of pumps, future studies should implement similar maximum burst points across conditions or utilise cognitive models without a scaling property.

It is also worth bearing in mind that the current version of the BART included forced choice trials to guide participants towards the optimising number of pumps in each phase. This manipulation was systematic; all participants followed the same instructions in the same trials throughout the experiment. Reassuringly, modelling only free choice or both free and forced-choice trials resulted in similar STL(-D) parameters estimates, suggesting that the inclusion of forced-choice trials did not obscure our results. Nevertheless, it would be reassuring to see converging results from other BART studies.

The neural basis of learning bias

Similarly to learning (Niv et al., 2012; Schultz, 1997; 2016; Frank et al., 2004; 2007; 2009), learning bias has been associated with dopaminergic and frontal cortical structures. Specifically, Lefebvre and colleagues (2107) found that higher bias was linked to increased reward prediction error signalling in the ventral striatum and ventromedial prefrontal cortex (vmPFC). Additionally, Sojitra et al. (2018) revealed that a polymorphism in the DARP-23 dopaminergic gene was associated with learning imbalance. Similarly, van den Bos et al. (2012) reported that the age-related reduction in negative feedback processing was related to increased connectivity between the striatum and medial prefrontal cortex. As in reward learning, frontal-subcortical connectivity (Moutsiana et al., 2015) as well as activity in the striatum and vmPFC (Kuzmanovic et al., 2016) were found to underlie the optimism bias in belief updating.

Given the high degree of similarity between cortical structures of a late reward learning system (Fouragnan et al., 2015; 2018) and those underlying dopamine-mediated reward learning (O'Doherty et al., 2001; Schultz et al., 1997; 2016) and optimism bias (Sharot et al., 2011; 2012), it is possible that the late system or interaction patterns across the early and late systems (Fouragnan et al., 2015) mediates learning bias. The latter possibility is further substantiated by the mounting evidence indicating prominent structures of the early system, such as the anterior cingulate cortex (ACC) or the thalamus (Fouragnan et al., 2015), in regulating reward learning (Behrens et al., 2007; Chakroun et al. 2020; Yu & Dayan, 2005; 2009). Accordingly, Sharot and colleagues (2007) indicated the ACC, which has strong reciprocal connections with the noradrenergic locus coeruleus (Briand et al., 2007; Joshi & Gold, 2020), in meditating the optimism bias in belief updating. Similarly, the thalamus, which was found to moderate the interaction between and early and late systems (Fouragnan et al., 2015) and plays a crucial role in avoidance learning (Kerns et al., 2004; Minamimoto et al., 2005; Seifert et al., 2011), has also been implicated in the processing of optimism bias (Kuzmanic et al., 2016). Moreover, recent work from our lab (Ban, 2024) suggests that the early system is linked to uncertainty processing as well as the locus-coeruleus noradrenergic system (LC-NA), which implies the potential modulatory role of these networks in partly generating learning bias. Even though we collected EEG data during the experiment, the paradigm was not originally designed for exploring the neural signatures of learning bias. As such, future research is needed to clarify how learning bias may be linked to neurotransmitter networks or structures of the early and late systems.

Conclusion

We provide evidence for a maladaptive learning bias in sequential decision making that is contingent on increased levels of uncertainty. Additionally, we found a consistent positive association between the degree of learning bias and risk-taking propensity, implying that the relative difference in learning from desirable and undesirable outcomes may generally guide risky behaviour. Future studies investigating the neural underpinnings of learning bias could investigate how reward learning is implemented in human frontal-subcortical networks and may be modulated by different neurotransmitter systems. Given the compromised reward learning processes associated with various neuropathological conditions (e.g., depression and anxiety disorders, Parkinson's disease, etc.) and the popularity of the BART in clinical research, learning bias could be investigated as an easy-to-derive measure for quantifying the severity of dysfunction.

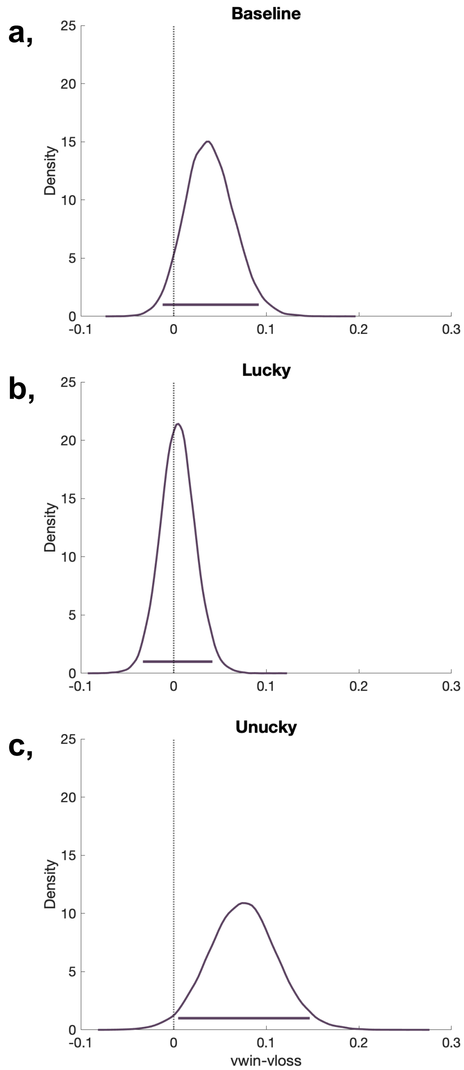
Supplementary methods & results

Sample size calculations

To test whether a learning bias (calculated as the normalised difference in differential learning rates; Eq.5) characterises behaviour in the BART, we planned to carry out undirected *t*-tests for each phase. We conducted an a priori power analysis in G*Power (version, ref) for sample size estimation. Our effect size was determined to be .5 based on a calculation including a null hypothesis mean of 0, an alternative hypothesis mean of .2, and a standard deviation of .4. We selected a conservative value of .2 for the difference in learning bias means, which was previously indicated to lie near .4 (Palminteri et al., 2017). Given the lack of existing results indicating the standard deviation of normalised learning rate bias in a full sample of participants, we opted to estimate our effect size based on the conservative estimate of SD = .4. With a significance criterion of α = .05 and power = .9, the minimum sample size required for our determined effect size was N = 44. Our obtained sample size of N = 50 should thus be appropriate for testing our central hypothesis.

Bayesian analysis of learning bias

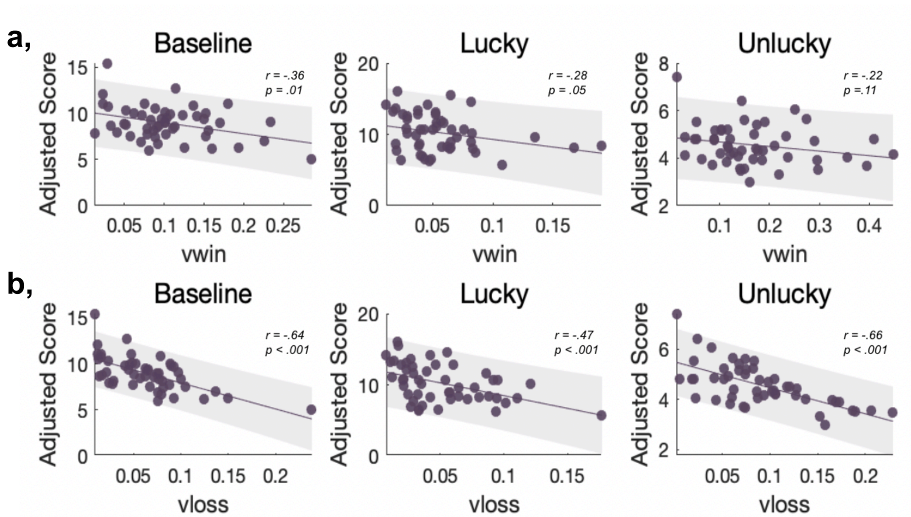
To confirm our results regarding the presence of learning bias in each experimental condition (Fig.6a), we compared group-level estimates of *vwin* and *vloss* from STL-D within the Bayesian framework. Specifically, we calculated the 95% highest density intervals (HDIs) to assess the group-level difference in learning rates, quantified by subtracting the group-level estimates of *vloss* from the group-level estimates of *vwin*. We considered the difference in the group-level learning rates to be credible if the 95% HDIs did not contain zero (Kruschke, 2014). We carried out this analysis separately for each experimental phase. We found the learning rate difference to be credible in the unlucky (95% HDI = [.01, .15], *M* = .07, *SD* = 0.04), but not in the baseline (95% HDI = [-.01, .09], *M* = .04, *SD* = 0.03) or lucky (95% HDI = [-.03, .04], *M* = .005, *SD* = 0.02) phases (Supplementary Figure 1). These results further indicate that the presence of learning bias is contingent on the level of uncertainty in the BART, with bias emerging under the condition with the highest level of uncertainty.



Supplementary Figure 1. Bayesian analysis of learning rate bias. Distribution for the difference in the group-level estimates of *vwin* and *vloss* is shown for the baseline (**a**), lucky (**b**), and unlucky (**c**) phases. The horizontal line within each distribution represents the 95% HDI. The difference in learning rates was credible in the unlucky (the vertical dotted line at 0 does not cross the bar reflecting the 95% HDI), but not in the baseline or lucky phase.

Learning rates and risk-taking propensity

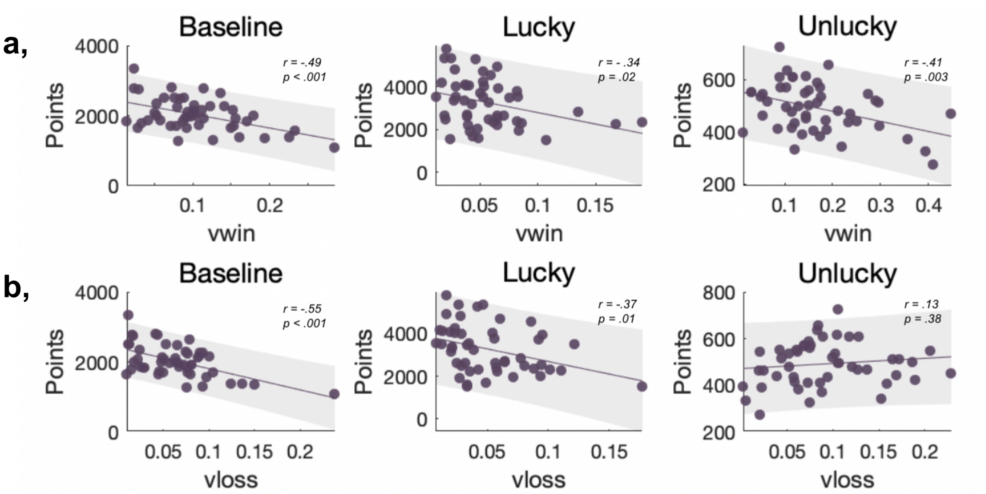
To evaluate the degree of association between individual-level learning rates *vwin* as well as *vloss* and risk-taking propensity, we employed across-participant Pearson's correlations. We utilised the adjusted score (mean number of pumps across unexploded balloons) as a proxy for risk-taking propensity. As Supplementary Figure 2 shows, the adjusted score significantly and negatively correlated with *vloss* in all phases (baseline: $r(48) = -.64$, $p < .001$, lucky: $r(48) = -.47$, $p < .001$, unlucky: $r(48) = -.66$, $p < .001$). Similarly, we found a significant negative correlation between the adjusted score and *vwin* in the baseline ($r(48) = -.36$, $p = .01$) and lucky ($r(48) = -.28$, $p = .045$) phases, but not in the unlucky phase ($r(48) = -.22$, $p = .11$). These results are consistent with previous findings that indicated a negative association between risk-taking propensity and learning rates (Niv et al., 2012; Palminteri et al., 2017), suggesting that elevated learning rates are linked to more risk-averse behaviour.



Supplementary Figure 2. Learning rates and risk-taking propensity.
Across-participant Pearson's correlation between individual-level parameter estimates for learning from wins *vwin* (**a**) as well as learning from losses *vloss* (**b**) from the STL-D model and the adjusted score are shown separately for each phase of the experiment. The Pearson's correlation coefficient and its corresponding *p*-value are shown on the top right of each graph. Data were aggregated across corresponding phases of Orders 1 and 2.

Learning rates and performance

To examine how learning rates *vwin* as well as *vloss* are linked to performance in each experimental phase, we utilised across-participant Pearson's correlations. We used the number of points earned in each experimental phase as a proxy for performance. As Supplementary Figure 3 depicts, we found a negative link between performance and *vwin* in all conditions (baseline: $r(48) = -.49$, $p < .001$, lucky: $r(48) = -.34$, $p = .002$, unlucky: $r(48) = -.41$, $p = .003$). Similarly, performance and *vloss* were negatively correlated in the baseline ($r(48) = -.55$, $p < .001$) and lucky ($r(48) = -.37$, $p = .01$) phases, whilst this association was not significant in the unlucky phase ($r(48) = .13$, $p = .38$). The negative correlation in the baseline and lucky phases implies that the more weight participants attributed to recent feedback, the more they overestimated environmental fluctuations, which in turn resulted in worse performance. On the other hand, in the unlucky phase, it is adaptive to learn predominantly from negative feedback, which explains the lack of a negative association between performance and *vloss*.



Supplementary Figure 3. Learning rates and performance. Across-participant Pearson's correlation between points earned and individual-level parameter estimates for learning from wins *vwin* (**a**) as well as learning from losses *vloss* (**b**) from the STL-D model are shown separately for each phase of the experiment. The Pearson's correlation coefficient and its corresponding *p*-value are shown on the top right of each graph. Data were aggregated across corresponding phases of Orders 1 and 2.

References

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. Annals of the Institute of Statistical Mathematics, 30(1), 9-14. doi:10.1007/BF02480194

Aklin, W. M., Lejuez, C. W., Zvolensky, M. J., Kahler, C. W., & Gwadz, M. (2005). Evaluation of behavioral measures of risk taking propensity with inner city adolescents. Behav Res Ther, 43(2), 215-228. doi:10.1016/j.brat.2003.12.007

Ban, K. (2024). *On the computational and neural characterisation of reward learning behaviour.* (PhD thesis). University of Glasgow, Enlighten theses. Retrieved from https://theses.gla.ac.uk/id/eprint/84313

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. Nat Neurosci, 10(9), 1214-1221. doi:10.1038/nn1954

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. Nat Neurosci, 10(9), 1214-1221. doi:10.1038/nn1954

Briand, L. A., Gritton, H., Howe, W. M., Young, D. A., & Sarter, M. (2007). Modulators in concert for cognition: modulator interactions in the prefrontal cortex. Prog Neurobiol, 83(2), 69-91. doi:10.1016/j.pneurobio.2007.06.007

Browning, M., Behrens, T. E., Jocham, G., O'Reillly, J. X., & Bishop, S. J. (2015). Anxious Individuals Have Difficulty Learning the Causal Statistics of Aversive Environments. Biological Psychiatry, 77(9), 47s-48s. Retrieved from <Go to ISI>://WOS:000352207500123

Carver, C. S., Scheier, M. F., & Segerstrom, S. C. (2010). Optimism. Clinical Psychology Review, 30(7), 879-889. doi:10.1016/j.cpr.2010.01.006

Cazé, R. D., & van der Meer, M. A. A. (2013). Adaptive properties of differential learning rates for positive and negative outcomes. Biological Cybernetics, 107(6), 711-719. doi:10.1007/s00422-013-0571-5

Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F., & Peters, J. (2020). Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. Elife, 9. doi:10.7554/eLife.51260

Corr, P. J. (2004). Reinforcement sensitivity theory and personality. Neurosci Biobehav Rev, 28(3), 317-332. doi:10.1016/j.neubiorev.2004.01.005

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. Neural Networks, 15(4-6), 603-616. doi:Pii S0893-6080(02)00052-7. doi 10.1016/S0893-6080(02)00052-7

den Ouden, Hanneke E. M., Daw, Nathaniel D., Fernandez, G., Elshout, Joris A., Rijpkema, M., Hoogman, M., . . . Cools, R. (2013). Dissociable Effects of Dopamine and Serotonin on Reversal Learning. Neuron (Cambridge, Mass.), 80(4), 1090-1100. doi:10.1016/j.neuron.2013.08.030

Éltető, N., Janacsek, K., Kóbor, A., Takács, A., Tóth-Fáber, E., & Németh, D. (2019). Do adolescents take more risks? Not when facing a novel uncertain situation. Cognitive Development, 50, 105-117. doi:10.1016/j.cogdev.2019.03.002

Fein, G., & Chang, M. (2008). Smaller feedback ERN amplitudes during the BART are associated with a greater family history density of alcohol problems in treatment-naive alcoholics. Drug and Alcohol Dependence, 92(1-3), 141-148. doi:10.1016/j.drugalcdep.2007.07.017

Fouragnan, E., Retzler, C., Mullinger, K., & Philiastides, M. G. (2015). Two spatiotemporally distinct value systems shape reward-based learning in the human brain. Nat Commun, 6, 8107. doi:10.1038/ncomms9107

Fouragnan, E., Retzler, C., Mullinger, K., & Philiastides, M. G. (2015). Two spatiotemporally distinct value systems shape reward-based learning in the human brain. Nat Commun, 6, 8107. doi:10.1038/ncomms9107

Fouragnan, E., Retzler, C., & Philiastides, M. G. (2018). Separate neural representations of prediction error valence and surprise: Evidence from an fMRI meta-analysis. Hum Brain Mapp, 39(7), 2887-2906. doi:10.1002/hbm.24047

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat Neurosci, 12(8), 1062-1068. doi:10.1038/nn.2342

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proceedings of the National Academy of Sciences of the United States of America, 104(41), 16311-16316. doi:10.1073/pnas.0706111104

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. Science, 306(5703), 1940-1943. doi:10.1126/science.1102941

Geana, A., Barch, D. M., Gold, J. M., Carter, C. S., MacDonald, A. W., Ragland, J. D., . . . Frank, M. J. (2022). Using Computational Modeling to Capture Schizophrenia-Specific Reinforcement Learning Differences and Their Implications on Patient Classification. Biological Psychiatry-Cognitive Neuroscience and Neuroimaging, 7(10), 1035-1046. doi:10.1016/j.bpsc.2021.03.017

Gelman, A. (2013). Bayesian data analysis (Third ed.). Boca Raton, Florida: CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. Statistical science, 7(4), 457-472. doi:10.1214/ss/1177011136

Gray, J. A. (1975). Elements of a two-process theory of learning. London: Academic Press.

Harada, T. (2020). Learning From Success or Failure? - Positivity Biases Revisited. Frontiers in psychology, 11. doi:ARTN 162710.3389/fpsyg.2020.01627

Joshi, S., & Gold, J. I. (2020). Pupil Size as a Window on Neural Substrates of Cognition. Trends Cogn Sci, 24(6), 466-480. doi:10.1016/j.tics.2020.03.005

Kerns, J. G., Cohen, J. D., MacDonald, A. W., 3rd, Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. Science, 303(5660), 1023-1026. doi:10.1126/science.1089910

Kóbor, A., Takács, A., Janacsek, K., Németh, D., Honbolygó, F., & Csépe, V. (2015). Different strategies underlying uncertain decision making: higher executive performance is associated with enhanced feedback-related negativity. *Psychophysiology, 52*(3), 367-377. doi:10.1111/psyp.12331

Kóbor, A., Tóth-Fáber, E., Kardos, Z., Takács, Á., Éltető, N., Janacsek, K., . . . Nemeth, D. (2023). Deterministic and probabilistic regularities underlying risky choices are acquired in a changing decision context. Scientific reports, 13(1), 1127-1127. doi:10.1038/s41598-023-27642-z

Kruschke, J. K. (2014). Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan. Amsterdam: Academic Press.

Kuzmanovic, B., Jefferson, A., & Vogeley, K. (2016). The role of the neural reward circuitry in self-referential optimistic belief updates. Neuroimage, 133, 151-162. doi:10.1016/j.neuroimage.2016.02.014

Kuzmanovic, B., & Rigoux, L. (2017). Valence-Dependent Belief Updating: Computational Validation. Frontiers in psychology, 8, 1087-1087. doi:10.3389/fpsyg.2017.01087

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. Nature Human Behaviour, 1(4). doi:ARTN 006710.1038/s41562-017-0067

Lejuez, C. W., Aklin, W. M., Jones, H. A., Richards, J. B., Strong, D. R., Kahler, C. W., & Read, J. P. (2003). The Balloon Analogue Risk Task (BART) differentiates smokers and nonsmokers. Exp Clin Psychopharmacol, 11(1), 26-33. doi:10.1037//1064-1297.11.1.26

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., . . . Brown, R. A. (2002). Evaluation of a Behavioral Measure of Risk Taking: The Balloon Analogue Risk Task (BART). Journal of experimental psychology. Applied, 8(2), 75-84. doi:10.1037/1076-898X.8.2.75

Minamimoto, T., Hori, Y., & Kimura, M. (2005). Complementary process to response bias in the centromedian nucleus of the thalamus. Science, 308(5729), 1798-1801. doi:10.1126/science.1109154

Moutsiana, C., Charpentier, C. J., Garrett, N., Cohen, M. X., & Sharot, T. (2015). Human Frontal-Subcortical Circuit and Asymmetric Belief Updating. Journal of Neuroscience, 35(42), 14077-14085. doi:10.1523/Jneurosci.1120-15.2015

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. The Journal of neuroscience, 32(2), 551-562. doi:10.1523/JNEUROSCI.5498-10.2012

Niv, Y., Joel, D., Meilijson, I., & Ruppin, E. (2002). Evolution of reinforcement learning in uncertain environments: A simple explanation for complex foraging behaviors. Adaptive Behavior, 10(1), 5-24. doi:Doi 10.1177/10597123020101001

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. Nature neuroscience, 4(1), 95. doi:10.1038/82959

Palminteri, S. (2023). Choice-Confirmation Bias and Gradual Perseveration in Human Reinforcement Learning. Behavioral Neuroscience, 137(1), 78-88. doi:10.1037/bne0000541

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. PLoS computational biology, 13(8), e1005684-e1005684. doi:10.1371/journal.pcbi.1005684

Palminteri, S., & Pessiglione, M. (2017). Opponent Brain Systems for Reward and Punishment Learning: Causal Evidence From Drug and Lesion Studies in Humans. Decision Neuroscience: An Integrative Perspective, 291-303. doi:10.1016/B978-0-12-805308-9.00023-3

Pleskac, T. J. (2008). Decision Making and Learning While Taking Sequential Risks. Journal of experimental psychology. Learning, memory, and cognition, 34(1), 167-185. doi:10.1037/0278-7393.34.1.167

Radulescu, A., Daniel, R., & Niv, Y. (2016). The Effects of Aging on the Interaction Between Reinforcement Learning and Attention. Psychology and Aging, 31(7), 747-757. doi:10.1037/pag0000112

R Core Team. (2021). “R: A language and environment for statistical computing.” R Foundation for Statistical Computing, Vienna, Astria. https://R-project.org/.

Schmitz, F., Manske, K., Preckel, F., & Wilhelm, O. (2016). The Multiple Faces of Risk-Taking Scoring Alternatives for the Balloon-Analogue Risk Task. *European Journal of Psychological Assessment, 32*(1), 17-38. doi:10.1027/1015-5759/a000335

Schultz, W. (2016). Dopamine reward prediction-error signalling: a two-component response. Nature Reviews Neuroscience, 17(3), 183-195. doi:10.1038/nrn.2015.26

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. Science, 275(5306), 1593-1599. doi:10.1126/science.275.5306.1593

Seifert, S., von Cramon, D. Y., Imperati, D., Tittgemeyer, M., & Ullsperger, M. (2011). Thalamocingulate interactions in performance monitoring. J Neurosci, 31(9), 3375-3383. doi:10.1523/JNEUROSCI.6242-10.2011

Seymour, B., Daw, N., Dayan, P., Singer, T., & Dolan, R. (2007). Differential Encoding of Losses and Gains in the Human Striatum. The Journal of neuroscience, 27(18), 4826-4831. doi:10.1523/JNEUROSCI.0400-07.2007

Sharot, T., Guitart-Masip, M., Korn, Christoph W., Chowdhury, R., & Dolan, Raymond J. (2012). How Dopamine Enhances an Optimism Bias in Humans. Current biology, 22(16), 1477-1481. doi:10.1016/j.cub.2012.05.053

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. Nature neuroscience, 14(11), 1475-U1156. doi:10.1038/nn.2949

Sharot, T., Riccardi, A. M., Raio, C. M., & Phelps, E. A. (2007). Neural mechanisms mediating optimism bias. Nature, 450(7166), 102-+. doi:10.1038/nature06280

Shepperd, J. A., Klein, W. M. P., Waters, E. A., & Weinstein, N. D. (2013). Taking Stock of Unrealistic Optimism. Perspectives on Psychological Science, 8(4), 395-411. doi:10.1177/1745691613485247

Smith, R., Taylor, S., Stewart, J. L., Guinjoan, S. M., Ironside, M., Kirlic, N., . . . Paulus, M. P. (2022). Slower Learning Rates from Negative Outcomes in Substance Use Disorder over a 1-Year Period and Their Potential Predictive Utility. Computational psychiatry, 6(1), 117. doi:10.5334/cpsy.85

Sojitra, R. B., Lerner, I., Petok, J. R., & Gluck, M. A. (2018). Age affects reinforcement learning through dopamine-based learning imbalance and high decision noise-not through Parkinsonian mechanisms. *Neurobiology of Aging, 68*, 102-113. doi:10.1016/j.neurobiolaging.2018.04.006

Spiegelhalter, D. J., Best, N. G., Carlin, B. R., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B-Statistical Methodology, 64, 583-616. doi:Doi 10.1111/1467-9868.00353

Stan Developmental Team. (2019). “RStan: the R interface to Stan.” R package version 2.17.5, https://mc-stan.org/.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: an introduction (Second ed.). Cambridge, Massachusetts: The MIT Press.

Thorndike, E. L. (1898). Review of: Animal Intelligence: An Experimental Study of the Associative Processes in Animals. Psychological Review, 5(5), 551-553. doi:10.1037/h0067373

van den Bos, W., Cohen, M. X., Kahnt, T., & Crone, E. A. (2012). Striatum-Medial Prefrontal Cortex Connectivity Predicts Developmental Changes in Reinforcement Learning. Cerebral Cortex, 22(6), 1247-1255. doi:10.1093/cercor/bhr198

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2011). Cognitive model decomposition of the BART: Assessment and application. Journal of mathematical psychology, 55(1), 94-105. doi:10.1016/j.jmp.2010.08.010

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing, 27(5), 1413-1432. doi:10.1007/s11222-016-9696-4

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling Behavior in a Clinically Diagnostic Sequential Risk-Taking Task. Psychological Review, 112(4), 862-880. doi:10.1037/0033-295X.112.4.862

Weinstein, N. D. (1980). Unrealistic Optimism About Future Life Events. Journal of Personality and Social Psychology, 39(5), 806-820. doi:Doi 10.1037/0022-3514.39.5.806

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. Psychonomic Bulletin & Review, 12(3), 387-402. doi:Doi 10.3758/Bf03193783

Yechiam, E., Busemeyer, J. R., Stout, J. C., & Bechara, A. (2005). Research Article: Using Cognitive Models to Map Relations Between Neuropsychological Disorders and Human Decision-Making Deficits. Psychological science, 16(12), 973-978. doi:10.1111/j.1467-9280.2005.01646.x

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. Neuron, 46(4), 681-692. doi:10.1016/j.neuron.2005.04.026

Zhou, R., Myung, J. I., & Pitt, M. A. (2021). The scaled target learning model: Revisiting learning in the balloon analogue risk task. Cognitive psychology, 128, 101407-101407. doi:10.1016/j.cogpsych.2021.101407