Task-irrelevant variability in reward magnitude and feedback salience bias reinforcement learning

Hans Kirschner
Institute of Psychology
Otto-von-Guericke University
D-39106 Magdeburg, Germany
hans.kirschner@ovgu.de

Matthew R. Nassar
Department of Neuroscience
Brown University
Providence RI 02912-1821, USA
matthew_nassar@brown.edu

Markus Ullsperger
Institute of Psychology
Otto-von-Guericke University
D-39106 Magdeburg, Germany
markus.ullsperger@ovgu.de

Abstract

Optimal decision-making requires organisms to adaptively adjust its sensitivity to new information. While numerous studies demonstrate that humans can adaptively weight task-relevant information based on the stochasticity and volatility of the environment, less is known about the influence of task-irrelevant factors on choice behavior. Here, we used computational modeling and EEG - as a brain measure with high temporal resolution - to better understand mechanisms responsible for the influence of task-irrelevant variability in reward magnitude and feedback salience on probabilistic learning. Specifically, we investigated learning behavior in a variant of a probabilistic reversal learning task with different levels of noise, that introduced two types of task-irrelevant events: pay-out magnitudes were varied randomly and, occasionally, feedback presentation was enhanced by visual surprise. We found that participants' learning performance was biased by distinct effects of these task-irrelevant factors. On the computational level, we show that both factors modulated trial-by-trial learning rate dynamics. In the EEG, these learning rate dynamics were reflected in a feedback-locked centroparietal positivity that also predicted behavioral adaptations. These results were replicated in an independent sample using a version of the task with reduced levels of noise. Interestingly, higher sensitivity to task-irrelevant factors was only negatively related to overall task performance in the task with high level of noise. Collectively, these data help to clarify the impact of task-irrelevant factors on probabilistic learning and suggest that these factors have a counter-normative influence on trial-by-trial learning rate dynamics.

Keywords: probabilistic reversal learning, learning bias, EEG

Acknowledgements

The authors thank Brontë Graham and Christian Bauer for help with the data collection. This research was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ERC advanced grant agreement No 101018805 to MU).

1 Introduction

Optimal decision-making requires organisms to adaptively adjust its sensitivity to incoming information. In environments with changepoints and noise, optimal learning requires amplifying the influence of surprising information during change and minimizing it during stable phases. On a computational level, one class of models suggest that this weighting is achieved via dynamic learning rate adjustments according to the statistical content of new information[1,9]. Effects of adaptive learning rates on optimal behavior in volatile environments have been studied in AI, autonomous systems and robotics (also called "dry") reinforcement learning (RL) literature.[2,4] In the wet RL literature, numerous studies demonstrate that humans adaptively integrate new information into their beliefs by considering the stochasticity and volatility of the environment [1,9]. While the exact neural implementation of dynamic learning rate adjustments is yet to be established, there is research highlighting the importance of the locus coeruleus/norepinephrine system in facilitating adaptive learning rate adjustments [11,12]. Moreover, several studies on EEG correlates of learning suggest that a late, feedback-locked centroparietal positivity referred to as the P300, tracks learning rate adjustments [3,10].

Importantly, in complex environments additional factors can influence learning, some of which even might be irrelevant for upcoming decisions but difficult to ignore. Indeed, there is evidence, that task-irrelevant information such as randomly varying outcome magnitudes or visual surprise are tied to learning and thus bias future outcome expectations [6,7]. Interestingly, the problem of learning in the presence of distractors has also been studied in the dry RL literature.[8] Yet, the neuro-computational mechanisms underlying these learning biases are not fully understood.

Here, we used EEG as a brain measure with high temporal resolution and computational modeling to better understand mechanisms responsible for the influence of task-irrelevant variability in reward magnitude and feedback salience on probabilistic learning. Participants (n=28) performed a probabilistic reversal learning task with different levels of noise. On each trial, subjects made a choice to either gamble or avoid gambling on a probabilistic outcome, in response to a stimulus presented in the middle of the screen (see **Figure 1A**). We introduced two types of task-irrelevant events: pay-out magnitudes were varied randomly and, occasionally, feedback presentation was enhanced by visual surprise. Both events were completely decorrelated from trial types and outcomes. We hypothesized that participants are unable to ignore these task-irrelevant factors despite explicit knowledge of their irrelevance for task performance and tested this hypothesis by investigating the effect of task-irrelevant factors on trial-by-trial learning rate dynamics. As predicted, we found that task-irrelevant factors biased learning on a given trial and that stronger biases were associated with worse task performance. Moreover, our results revealed that the P300 relates to learning rate dynamics and predicts behavior adaptations. These results were replicated in an independent sample (n=19), which also demonstrated that the negative effect of learning biases on task performance depend on the level of noise in the environment.

Part of this work has been published in Kirschner et al., (2022)[6], which presents the EEG correlates of task-irrelevant factors in detail. The dynamic learning rate model, the EEG correlates of the learning rate, and the replication study are new contributions. The code of the RL models can be accessed here: https://github.com/HansKirschner/SaLe_Model

2 Results

In the task, to maximize financial earnings, participants had to learn the reward probabilities of three stimuli. On each trial, they could gamble on a stimulus and win or lose 10 or 80 points (translating to €0.10/€0.80) or choose to avoid gambling and observe what would have happened, without any financial consequences. During the task, reward probabilities of the stimuli could change unexpectedly. Moreover, we introduced two task- irrelevant factors a) visual surprise —on 20% of trials, the feedback background briefly flashed green (positive) or red (negative) instead of the standard black **(Figure 1B)**; and (b) random payout magnitudes—on 50% of trials, outcomes involved either 10 or 80 points **(Figure 1C)**. Participants' choices largely reflected the stimuli's reward probabilities, indicating successful task learning. (**Figure 2A**). In a first analysis, we aimed to determine whether both types of task irrelevant factors had an effect on participants' behavior. We found that both types of task-irrelevant events prolonged RTs on subsequent trials ($\beta_{post\ visual\ surprise}$ = 0.04, 95% credible interval = [0.030, 0.068], ΔRT = ~23 ms; $\beta_{post\ magnitude}$ = 0.019, credible interval = [0.006, 0.031], ΔRT = ~12 ms). Moreover, the reward magnitudes manipulation affected participants' choice probabilities. Specifically, participants were more likely to gamble on a stimulus after receiving a high vs. a low positive reward. Vice versa, they were more likely pass gambling on a stimulus after high vs. low losses. ($\beta_{post\ outcome\ x\ magnitude}$= -0.65 95% credible interval = [0.447, -0.847]; see **Figure 2B**).

Having established that both task-irrelevant factors influenced participants' behavior, we next sought to parse out the computational mechanisms of this observation. We began by fitting an ideal observer model for inference in the presence of changepoints[1,5] to the task environment. This approach allowed us to first derive a normative learning rate free of any contamination by task-irrelevant factors **(Figure 1E)**. Next, to capture trial-to-trial dynamics of subjective probability assessments, we fit the choice behavior from each participant with a reinforcement learning (RL) model that incrementally learned the probability of reward from feedback. The model updated expected reward probabilities for each trial with a dynamic learning rate. The model included a baseline coefficient capturing the average rate of learning as well as coefficients that allowed the learning rate to increase or decrease according to three factors: trial-to-trial learning rate from the ideal observer model, reward magnitude, and visual surprise (the latter two reflecting possible counter-normative influences on learning rate). Moreover, the model considered that the expected reward probability on a given trial was influenced by learned reward probabilities, along with a parameter capturing fixed biases toward gambling irrespective of the expected probability of reward. Coefficients that described the effects of ideal learning rate and reward magnitude on learning rate were positive across participants, while the effect of visual surprise was centered around zero across participants (**Figure 2C**; ideal learning rate: two-tailed $t_{(27)}$ = 3.17, p = 0.003, d = 0.61, 95% confidence interval = [0.11, 0.49]; reward magnitude: $t_{(27)}$ = 4.67, p >.001, d = 0.89, confidence interval = [0.35 ,0.89]; visual surprise: $t_{(27)}$ = -1.05, p = 0.30, d = -0.20, confidence interval = [-0.27, 0.08]). In other words, participants were more responsive to feedback that was provided during a period with a high normative learning rate and with high reward magnitudes. Thus, the ideal learning rate and reward magnitudes scaled the extent to which reward prediction errors (RPEs) were used to adjust subsequent behavior. To better understand how learning biases relate to overall task performance, we examined how performance in the task related to model parameters from our fits. To do so, we regressed task performance onto an explanatory matrix containing our model parameter estimates (**Figure 2D**). The results of this analysis revealed that counter-normative influences on learning were negatively predicting overall task performance ($\beta_{reward\ magnitude}$ = -.624, credible interval = [-1.008, -.220]); ($\beta_{visual\ surprise}$ = -.749, credible interval = [-1.126, -.359]).

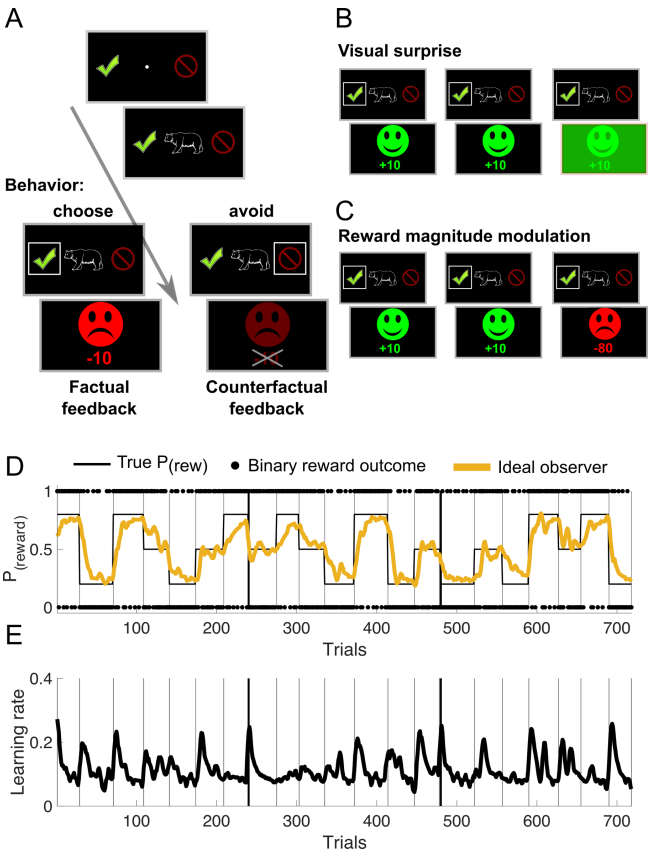

Figure 1. Task and ideal observer model. (A) Schematic of the probabilistic reversal learning task. On each trial, a fixation dot and choice options were presented for 300 –700 ms. Next, the stimulus was presented for up to 1700 ms. During this time participants had to decide, if they wanted to gamble on the stimulus or not. After subject's decisions their choice was highlighted for 350 ms. Finally, depending on participants choice, either factual or counterfactual feedback was presented for 750 ms. **(B)** On 20% of the trials, the color of the feedback background changed from black to a feedback-matching color (i.e. red or green), introducing task irrelevant visual surprise. **(C)** A second manipulation in the task focused on the reward magnitudes. Here, magnitudes randomly varied between 10 and 80 points. **(D)** Plot showing model predictions. Reward probabilities reversed occasionally to require learning (solid black line). Bold vertical black lines indicate a change of stimuli and thin vertical black lines indicate reversals. Binary outcomes (black dots), which were governed by the underlying reward probabilities, were used by an ideal observer model for inference in the presence of changepoints[1,5] (yellow line). **(E)** Plot showing the ideal learning rate derived from the ideal observer model. In this model the learning rate for a new observation is mathematically defined and depends on the uncertainty about the underlying (reward) distribution and the likelihood of a change-point. It is noteworthy that normative learning is typically larger after changepoints. **(A-C)** adopted from [6]. Note that stimuli were sorted in D&E for ease of visualization but presented interleaved in the task.

Next, we explored the counter-normative influence of reward magnitude on learning rate. Here, we split the sample via a median-spilt on the individual reward magnitude parameter estimates. This analysis revealed higher and less dynamic learning rates for participants with higher reward magnitude coefficients (**Figure 3A**). To investigate the neural correlates of the learning rate, we regressed feedback-locked EEG data collected

simultaneously with task performance onto the trial-by-trial learning rate derived from the RL model. Regression weights were aggregated across subjects to create a map of t-statistics (**Figure 3B**), and spatiotemporal clusters of electrode/timepoints exceeding a cluster-forming threshold were tested against a permutation distribution of cluster mass to spatially and temporally organized fluctuations in voltage that related to learning rate dynamics. This procedure yielded a large cluster of centroparietal positive coefficients spanning 320–750ms post feedback, matching the timing, direction and topography of the canonical P300 response. Next, we used multivariate pattern analysis on feedback locked EEG activity of the whole scalp and trained a support vector machine to predict behavioral switches (changes in choice behavior on the next encounter of stimuli with the same identity). Cluster-based permutation analyses revealed a large cluster of time points at which the decoding of future switches was significantly greater than chance level (**Figure 3C**). Projecting classifier weights onto EEG sensor space at the maximal individual decoding accuracy (mean ± s.e.m.; 503 ± 38.08ms) revealed a topographic match with the P300 which occurs in this time range (**Figure 3C, inset**). Finally, temporal generalization analyses demonstrate that the classifier predicting behavior switches generalizes within the P300 latency (**Figure 3D**).

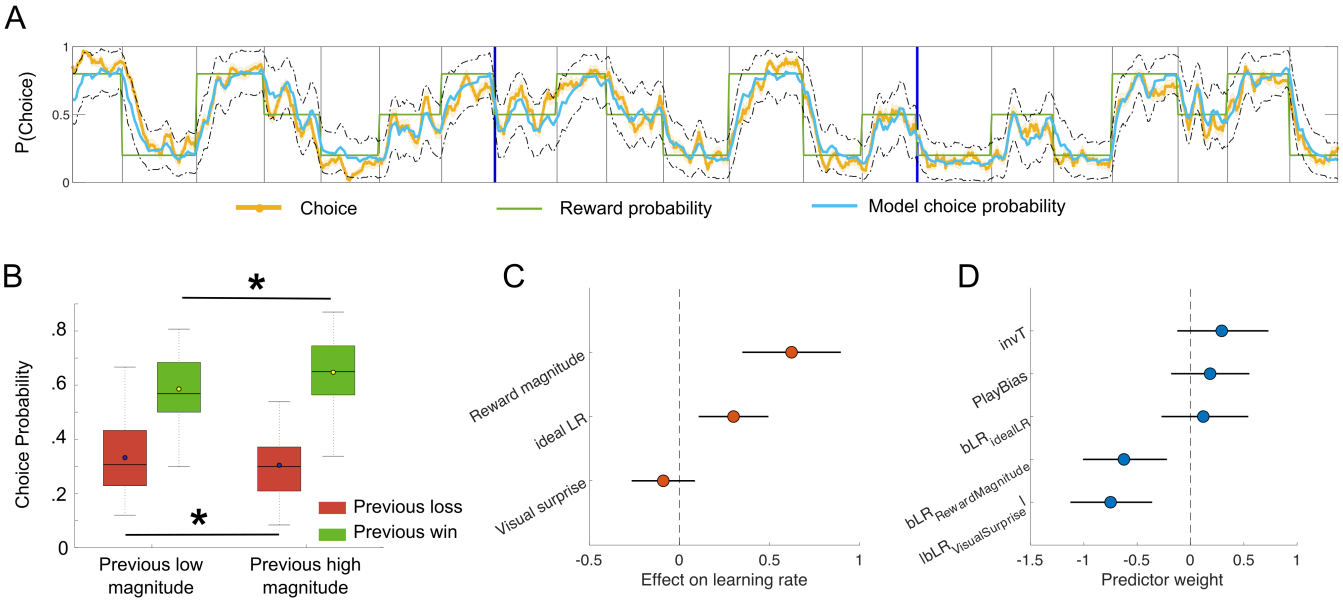

Figure 2. Task-irrelevant factors affect choice behavior and learning rates. (A) Modelled and choice behavior, stretched out for all stimuli. Note that in the task the different animal stimuli were presented in an intermixed and randomized fashion, but this visualization allows to see that participants' choices followed the reward probabilities of the stimuli. Data plots are smoothed with a running average (+/- 2 trials). Ground truth corresponds to the reward probability of the respective stimuli (good: 80%; neutral: 50%; bad: 20%). Dashed black lines represent 95% confidence intervals derived from 1000 simulated agents with parameters that were best fit to participants in each group. **(B)** Raw value splits for the effect of previous reward magnitude and outcome on the choice probability on a given trial. **(C)** Mean maximum likelihood estimates and 95% confidence intervals of parameters affecting learning rate in the model **(D)** Regression coefficients and 95% credible intervals (points and lines; sorted by value) stipulating the contribution of each model parameter estimate to overall participants task performance (i.e., scored points in the task). * = significant difference.

2.1 Replication of the main results in a follow-up experiment

Our initial findings suggested that task-irrelevant variability in reward magnitude and, partly, feedback salience bias probabilistic learning. However, the relatively large proportion of neutral reward probabilities may have facilitated this bias, because of the increased relative uncertainty about the stimuli value. To address this issue, we conducted a follow-up experiment in an independent sample of healthy young adults (n = 19). In the new learning task, the reward probabilities for the three stimuli only varied between 20% and 80% reward probability. We replicated all effects of the initial study. The only difference that we observed, was that variability in learning bias was no longer related to overall task performance ($\beta_{reward\ magnitude}$ = -.231, credible interval = [-0.676, 0.241]; $\beta_{visual\ surprise}$ = -.571, credible interval = [-1.332, 0.385]).

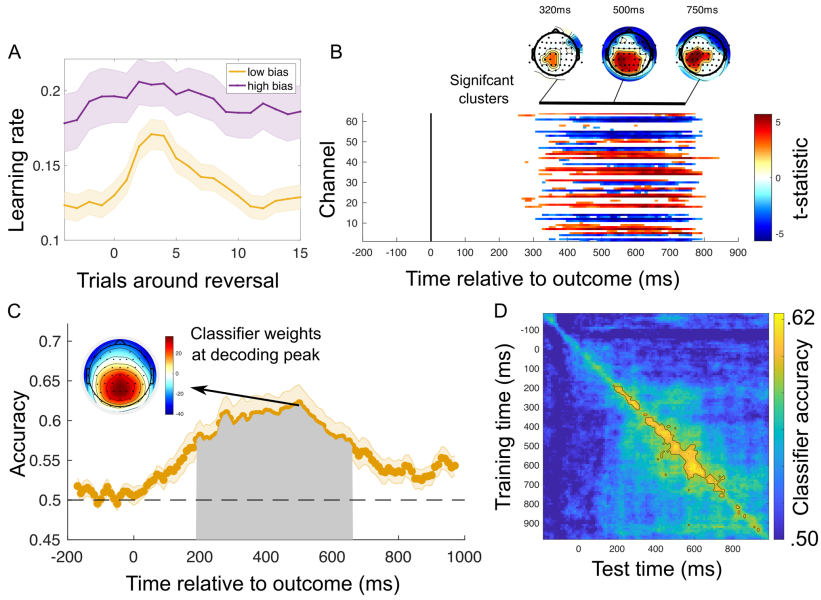

Figure 3. Outcome-locked P300 reflects Learning rate dynamics predicts behavioral adaptations. (A) Learning rate around reversals for high and low learning bias based on median split on reward magnitude coefficients. **(B)** T-statistic map for clusters that survived multiple comparisons correction via permutation testing for learning rate contrast along with corresponding topoplots. **(C)** Mean accuracy of decoding of behavioral switches. Grey area indicates the cluster in which decoding accuracy significantly surpassed chance level after correction for multiple comparisons via permutation testing. Topoplot shows projected classifier weights onto EEG sensor space at the maximal individual decoding accuracy. **(D)** Temporal generalization of decoding performance. Contours indicate time point pairs where the generalization above chance after correction for multiple comparisons via permutation testing.

3 Conclusion

In this study, we combined computational modeling and EEG as a brain measure with high temporal resolution to better understand mechanisms responsible for the influence of task-irrelevant variability in reward magnitude and feedback salience on probabilistic learning in two independent samples. Collectively, these data demonstrate a counter-normative influence of reward magnitude on learning rates across participants. Moreover, counter-normative influences on learning are negatively predicting overall task performance depending on the level of probabilistic noise in the task. In the EEG, learning rate dynamics are reflected in a centroparietal positivity that also predicts behavioral adaptations.

4 Literature

1. Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. *Nat Neurosci* 10:1214-21.
2. Degris T, Javed K, Sharifnassab A, Liu Y, Sutton R (2024) Step-size optimization for continual learning. *arXiv preprint arXiv:240117401*.
3. Fischer AG, Ullsperger M (2013) Real and fictive outcomes are processed differently but converge on a common adaptive mechanism. *Neuron* 79:1243-55.
4. Fosong E, Rahman A, Carlucho I, Albrecht SV (2023) Learning complex teamwork tasks using a given sub-task decomposition. *arXiv preprint arXiv:230204944*.
5. Jang AI, Nassar MR, Dillon DG, Frank MJ (2019) Positive reward prediction errors during decision-making strengthen memory encoding. *Nature Human Behaviour* 3:719-32.
6. Kirschner H, Fischer AG, Ullsperger M (2022) Feedback-related eeg dynamics separately reflect decision parameters, biases, and future choices. *Neuroimage* 259:119437.
7. McGuire JT, Nassar MR, Gold JI, Kable JW (2014) Functionally dissociable influences on learning rate in a dynamic environment. *Neuron* 84:870-81.
8. McInroe T, Schäfer L, Albrecht SV (2022) Multi-horizon representations with hierarchical forward models for reinforcement learning. *arXiv preprint arXiv:220611396*.
9. Nassar MR, Wilson RC, Heasly B, Gold JI (2010) An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *J Neurosci* 30:12366-78.
10. Nassar MR, Bruckner R, Frank MJ (2019) Statistical context dictates the relationship between feedback-related eeg signals and learning. *Elife* 8.
11. Silvetti M, Vassena E, Abrahamse E, Verguts T (2018) Dorsal anterior cingulate-brainstem ensemble as a reinforcement meta-learner. *PLoS Comput Biol* 14:e1006370.
12. Yu LQ, Wilson RC, Nassar MR (2021) Adaptive learning is structure learning in time. *Neurosci Biobehav Rev* 128:270-81.