Computational Imprecision in Value-Based Learning and Decision-Making: Investigating the Potential Role of the Noradrenergic Locus Coeruleus System

Aleksic, Mladena

Computational Imprecision in Value-Based Learning and Decision-Making: Investigating the Potential Role of the Noradrenergic Locus Coeruleus System

Mladena Aleksic

s2146711

Faculty of Social and Behavioural Sciences, Leiden University

m.aleksic@umail.leidenuniv.nl

Master Thesis

supervised by Dr. Franz Wurm

August 26, 2023

Table of Contents

Abstract

Human decision-making in the context of value-based learning considerably deviates from the premise to always maximise short-term reward. This sort of behavioural variability has been attributed to exploration during the choice process and has been linked to the locus coeruleus-noradrenaline (LC-NA) system. A recent computational account implementing a "noisy" reinforcement learning (RL) model provides evidence that computational imprecision during the learning process contributes to behavioural variability in decision-making as well. In the present study, the role of the LC-NA system in modulating learning imprecision has been investigated, using a multi-modal approach including the newly developed noisy RL model, a double-blind pharmacological manipulation, and EEG. The responses of thirty participants in a decision-making task were investigated during two experimental sessions in which either atomoxetine, a NA reuptake inhibitor, or placebo was administered, and the model fit of the noisy RL model was tested against the classical RL model without learning imprecision parameter. Contrary to our expectations, we did not find evidence for a modulatory role of the LC-NA system of learning imprecision. However, in line with our expectations, increased NA levels did not impact exploratory behaviour either. Crucially, the noisy RL model outperformed the classical RL model assuming exact learning and learning imprecision led to a significant number of non-maximising decisions contributing to behaviour variability. Moreover, the exploratory EEG analyses on potential underlying mechanisms of learning imprecision suggest learning noise to emerge broadly during feedback processing, rather than within a certain time window. The current study provides further evidence for the importance of computational imprecision during value-based learning.

Computational Imprecision in Value-Based Learning and Decision-Making: Investigating the Potential Role of the Noradrenergic Locus Coeruleus System

In a world of ever-changing possibilities to satisfy one's needs, human decision-making has been found to be subject to variability: Decisions for or against an option fluctuate over time and do not always align with the highest expected reward (Sutton & Barto, 2018). A growing body of evidence suggests that noradrenaline (NA) released in the brainstem nucleus locus coeruleus (LC) plays a pivotal role in modulating this kind of behavioural variability (Aston-Jones & Cohen, 2005), however, the exact role and underlying mechanisms contributing to such variability remain to be determined. Theoretical work in the domain of reinforcement learning (RL) and decision-making assigns the role of regulating explorative versus exploitative behaviour to the LC-NA system, further suggesting that behavioural variability results from the need to explore lesser-known alternatives to mitigate uncertainty and potentially increase long-term reward. Recent theoretical and empirical work has introduced another source of behavioural variability, showing that learning imprecision during the learning process (i.e., linking an action to a certain reward) contributes to behavioural variability as well (Findling et al., 2019). The current study aims to investigate the role of the LC-NA system in regulating these two sources of behavioural variability in human learning and decision-making. Specifically, we hypothesise that the activity of the LC-NA system drives behavioural variability by modulating the level of learning imprecision, rather than regulating the "exploration-exploitation" trade-off. With the current research, we seek a direct test of the involvement of the LC-NA system in modulating learning imprecision during value-based learning by using a pharmacological manipulation to increase central NA levels. Moreover, we wish to investigate the dynamics of learning imprecision on the neural level using electroencephalography (EEG) due to its high temporal resolution.

Reinforcement learning (RL) is a computational framework for understanding how humans learn complex behaviour through simple rewarding or punishing signals (Lindsay, 2021). One of the main premises of RL is that humans as well as artificial agents aim at maximising reward in the long run (Sutton & Barto, 2018). However, our world presents challenges to achieving that goal, as immediate rewards and long-term values of actions often are unknown or subject to volatility (i.e., action-outcome contingencies change over time). One approach within RL is value-based or reward-guided learning, which involves trial-and-error learning to identify the most rewarding options. This approach follows a sequential process of learning and decision-making. Initially, the values of possible actions are estimated, providing a basis for subsequent decisions in the action selection process. Feedback on the chosen action is received, which in turn allows for an update of the action values. The updated action values then guide future decisions, resulting in an iterative learning and decision-making process.

Despite the simplicity of these steps, a key challenge in RL lies in effectively utilising the estimated values to make decisions that maximise outcome, as seen in the exploration-exploitation dilemma. While exploitation involves choosing the option with the highest estimated value to maximise short-term reward (so-called greedy decisions), exploration is necessary to potentially discover better alternatives to maximise long-term outcome by venturing beyond the option with the highest expected value (non-greedy decisions). Thus, balancing reward maximisation vs. information seeking is crucial for adaptive behaviour and optimal decision-making. Reportedly, humans make a substantial number of decisions which fail to maximise immediate reward, especially in uncertain and volatile environments (Jepma et al., 2020). Many RL and decision-making theories have predominantly attributed this phenomenon to exploration alone (Daw et al., 2006; Dayan & Daw, 2008), overlooking other potential factors contributing to non-greedy decisions.

Recent research by Findling et al. (2019) has shed light on such an additional mechanism contributing to non-greedy decisions, namely imprecision during the learning process. While classical RL models rely on precise reward prediction errors (RPE; obtained reward minus expected reward) for updating action values, the newly introduced “noisy RL model” incorporates noise along with the RPE in the updating process. This model suggests that decision-making is not solely guided by precise action values but also by noisy action values that are corrupted by learning noise. Consequently, the noisy RL model additionally entails random variability on the learning level. In contrast, exploration is accounted for at the choice level through random variability during the action selection process (i.e., choice stochasticity), where the degree of exploration influences the extent to which decisions are driven by estimated action values. The idea of human-decision making being corrupted by imprecision founds support from earlier studies, in which human perceptual decision-making was shown to be affected by inference noise during evidence accumulation (Drugowitsch et al., 2016; Findling & Wyart, 2021; Wyart & Koechlin, 2016).

To test if noise influences value-based decision making, Findling et al. (2019) compared the performance of their noisy RL model with the performance of the “exact” RL model in a canonical two-armed bandit task. Their findings indicated that the noisy RL model explained human decision-making better, with decisions labelled as non-greedy under the exact RL model often being classified as greedy under the noisy RL model. This suggests that learning imprecision might play a crucial role in driving behavioural variability. Additionally, the study revealed that the effect of learning imprecision was more pronounced in task conditions where exploration was unnecessary (i.e., during complete feedback conditions where both the chosen and forgone outcome was revealed) compared to conditions where exploration was required (i.e., in partial feedback conditions where only the chosen outcome is known). Overall, this emerging line of research highlights the contribution of learning imprecision, alongside the trade-off between exploitation and exploration, to behavioural variability in decision-making. Understanding the neurobiological mechanisms underlying such variability, particularly in the context of value-based learning, has been the focus of further investigation.

The locus coeruleus-noradrenaline (LC-NA) system is a prominent candidate for modulating behavioural variability during value-based decision-making and learning, given its neuromodulatory role and involvement in arousal, attention, learning, and memory (for reviews, see Sara, 2009; Sara & Bouret, 2012; Thiele & Bellgrove, 2018). While the seminal work of Schultz, Dayan, and Montague (1997) has linked dopaminergic activity in the ventral tegmental area (VTA) and substantia nigra with the reward prediction error, the influential adaptive gain theory proposes that tonic LC-NA activity regulates the exploration-exploitation trade-off (Aston-Jones & Cohen, 2005). By modulating the sensitivity (i.e., gain) of cortical target neurons to sensory input, cognitive and behavioural performance is optimised across situations that require either exploration (high tonic NA levels) or exploitation (low tonic NA levels). Empirical findings on that account are conflicting, however. For instance, Jepma et al. (2010) found that increased tonic NA levels after administering reboxetine, a selective NA reuptake inhibitor, had no effect on exploratory behaviour during a four-choice gambling task. Conversely, another study using atomoxetine, a NA reuptake blocker, to increase extracellular NA levels in the cortex, found an effect of increased tonic NA levels on exploration, however, in the opposite direction as hypothesised: higher tonic NA levels were associated with less exploration in a gambling task (Warren et al., 2017). Considering the contradictory results regarding the relationship between the LC-NA system and exploratory behaviour, an alternative explanation for the underlying mechanism of behavioural variability has been proposed by Findling et al. (2019). Specifically, their pupillometry data revealed a connection between trial-to-trial fluctuations in pupil dilation and learning noise. Given that pupil diameter has been previously linked to tonic LC-NA activity (Rajkowski et al., 1994), these findings imply a potential involvement of the LC-NA system in the modulation of learning imprecision.

Event-related potentials (ERPs) have been widely used to identify the underlying neural mechanisms of cognitive processes. The two primary ERPs associated with reward and feedback processing are the feedback-related negativity (FRN; for reviews, see Kirsch et al., 2022; Sambrook & Goslin, 2015) and the P3 (for an overview, see Polich, 2020). The frontocentrally distributed negative-polarity FRN typically occurs 200 to 300 ms after feedback delivery and is told to stem from dopaminergic anterior cingulate cortex (ACC) activity (Holroyd & Coles, 2002). The more parietal and positive-polarity component, the P3, occurs 300 to 700 ms after feedback and has been linked to phasic LC-NA activity (Nieuwenhuis, 2011; Nieuwenhuis et al., 2005). The FRN and P3 have both been found to be sensitive to feedback valence (higher negativity for losses than for wins), reward magnitude (higher negativity for higher than for lower rewards), and the probability of reward (higher negativity for rare than for common rewards) with conflicting findings however (San Martín, 2012; Yeung & Sanfey, 2004). In particular, the FRN amplitude has been suggested to scale with a signed RPE, exhibiting a more negative deflection when the outcome is smaller or worse (negative RPE) than when the outcome is larger or better than expected (positive RPE; Holroyd & Coles, 2002; for review, see Kirsch et al., 2022). The P3 amplitude on the other hand has been linked to an unsigned RPE (Mars et al., 2008) and to unexpected events (e.g., the oddball paradigm; Duncan-Johnson & Donchin, 1977), which both reflect surprise. Moreover, current evidence points to the possibility of a two-step feedback evaluation process, in which the FRN reflects a first evaluation of primary feedback attributes such as valence, and the P3 a later in-depth evaluation of secondary feedback attributes such as magnitude and probability (Bernat et al., 2015). To date, no study has explored the effect of learning imprecision on the FRN and P3. Thus, investigating ERP data may help bridge the gap between underlying neural mechanisms, cognitive processes, and computational models of human reinforcement learning and decision-making.

The current study aims to examine the role of the LC-NA system in regulating computational precision in value-based learning using a pharmacological manipulation to increase central NA levels, by administering a single dose of 40mg of atomoxetine. Atomoxetine, a selective NA reuptake inhibitor, is commonly used in the treatment of Attention Deficit Hyperactivity Disorder (ADHD) symptoms (Clemow et al., 2017) and was found to alleviate ADHD symptoms by increasing NA levels in the prefrontal cortex (PFC; Bymaster et al., 2002). One goal of the study is to replicate some relevant behavioural and computational findings of the study of Findling et al. (2019) as depicted in Figure 1. More specifically, we want to demonstrate that the model accounting for computational noise by entailing a learning imprecision parameter (noisy RL model) explains the human data better than an exact RL model without learning imprecision parameter (H1a). Furthermore, we expect the fraction of non-greedy decisions explained by learning imprecision to be lower when counterfactual information about the decisions is available than when it is not (H1b). We further hypothesise that increased NA levels are associated with increased learning imprecision but not with increased choice stochasticity (H2). Moreover, we want to explore potential model-free correlates such as reward magnitude (H3a) and model-based correlates such as learning noise (H3b) of FRN and P3 activity, and test if the effect of the model-based variables on FRN and P3 amplitude are modulated by NA or the amount of feedback information available (H3c).

Methods

Participants

Thirty Participants (17 female) aged 19-29 years (*M* = 22.87, *SD* = 2.27) were recruited via Leiden University’s online recruitment tool (https://ul.sona-systems.com/), designated Facebook groups, and posters distributed on campus (see Appendix A for the recruitment flyer). The following inclusion criteria were applied: 18-35 years old, no history of neurological or psychiatric disorders, no use of psychotropic medication or active drug use, no contra-indications to atomoxetine, normal weight (BMI: 18.5-25), normal or corrected-to-normal vision, and no dreadlocks (due to the EEG measurement). Participants were screened by a medical doctor to ensure eligibility to take atomoxetine, excluding those with certain medical conditions or habits, such as cerebral diseases (migraine, epilepsy, head trauma), cardiovascular diseases, hypertension, gastro-intestinal diseases (e.g., chronic inflammation), liver or kidney problems, premature birth, glaucoma, possible pregnancy, breast-feeding, heart arrythmia, alcohol addiction, use of anti-depressants, and smoking more than five cigarettes per day. Participation in the whole study took around 7.5 hours and was compensated with € 110 and a performance-based bonus of € 0, 5, or 10 in each session (*M* = € 9.02, *SD* = 2.38). No participant had to be excluded despite one experiencing side effects (nausea) from taking atomoxetine. All participants gave consent; the study was approved by the Medical Ethics Committee Leiden The Hague Delft (METC LDD).

Procedure

Study Design

The study followed a pseudo-random, double-blind, 2 (treatment) x 2 (feedback) within-subject design. We utilised a methodological procedure similar to that of Findling et al.'s (2019) behavioural study and were provided with their original material. In contrast to the previous study, our design included the use of EEG to measure brain activity, a pharmacological manipulation for a more direct test of the involvement of the LC-NA system, and our task design did not entail “choice-free, cued-trials”, in which participants had to choose the highlighted options.

Treatment Condition: Atomoxetine vs. Placebo. To test the effect of increased central NA levels on learning imprecision and brain activity, participants were orally administered a single dose of 40 mg atomoxetine - a selective NA reuptake inhibitor - in one session and placebo (microcrystalline cellulose PH 102) in the other session, counterbalanced across participants and feedback condition. Experimenter and participants were blinded to the treatment condition, and the substances were delivered by the pharmacy in identical containers. Both sessions occurred at the same time of day, followed the same experimental procedure, and lasted approximately 210 minutes each, with exactly one week between sessions. The initial dose of 40mg is generally considered safe and has a low risk of side effects (Heil et al., 2002). After administration, atomoxetine reaches its maximum plasma concentration between 60 to 120 minutes with variable half-life duration between extensive (5.2 hours) and poor metabolisers (21.6 hours; Sauer et al., 2005).

Feedback Condition: Partial vs. Complete. During the bandit task, participants received feedback for each decision. In half of the blocks feedback was delivered completely and in the other half partially. In the partial feedback condition, participants only received feedback about the chosen option, while both outcomes (chosen and forgone) were presented in the complete feedback condition. The block-design was pseudo-randomised, with conditions alternating between blocks. Having these two conditions allows for distinguishing between the sources of non-greedy decisions. In the partial feedback condition, non-greedy choices can arise from two possible sources: learning imprecision and the need for exploration. In contrast, the need to explore is expected to be unnecessary in the complete feedback condition because all information is revealed to the participant.

Session Procedure

Participants were asked to refrain from consuming alcohol for 24 hours and from coffee for 3 hours prior to each session. They were advised to have a regular meal up to one hour before the start of the session and to only consume water in the final hour before their appointment. Upon arrival at the laboratory, participants read and signed the informed consent form. To establish baseline α-amylase levels, a known biomarker for central NA activity and release (for review, see Segal & Cahill, 2009), participants provided a first saliva sample five minutes before taking either the active substance or the placebo. We used the passive drool collection method (Beltzer et al., 2010) to collect saliva at three measurement points (-5 minutes, 85 minutes, and 195 minutes after pill administration). Following drug administration, the electrocardiogram (ECG) was set up for an offline measurement of the heart rate throughout the session, indicating central NA release and activity. Participants then completed a brief demographics questionnaire and the State-Trait Anxiety Inventory (STAI; Spielberger et al., 1983), since trait anxiety has been linked to tonic LC-NA activity (Howells et al., 2012), implying a potential influence on the effect of NA release on task performance. After the questionnaires, participants received instructions for the bandit task. Two training blocks of the bandit task were conducted in the laboratory room, and feedback on the performance was provided to ensure participants' comprehension of the experiment. Subsequently, the EEG was set up, and the distance to the screen was maintained at 70cm. A second saliva sample was collected from participants five minutes before beginning the main experiment, consisting of 12 blocks of the bandit task. Exactly 90 minutes after pill administration, participants started with the experiment in the laboratory room, with dimmed lights and closed doors. After half of the blocks, participants had the opportunity to take a break and have a small snack. Upon finishing the experiment, participants were asked to guess which pill they received that day (with 58.33% of participants providing correct answers) and provided the final saliva sample 195 minutes after pill administration.

Measures

Restless Bandit Task

Participants completed a restless two-armed bandit task during which they chose between two options and were instructed to maximise outcome. Each trial began with a fixation cross, followed by the presentation of two differently coloured shapes. Participants indicated their choice by pressing the F key with the left index finger for the left option, and the J key with the right index finger for the right option (Figure 2a). Subsequently, feedback of the outcome (ranging from 1 to 99 points) was presented. In total, participants did the task for 12 blocks à 56 trials each session, and for two training blocks à 48 trials in the beginning of the sessions. The shapes and colours of the stimuli were randomly selected from a predetermined set, ensuring that the combinations were unique between blocks and that colours did not repeat within a single block to maximise distinction between the options. Additionally, to introduce volatility into the decision-making environment, a random walk procedure (for details, see Findling et al., 2019) was employed for each block separately, to generate average outcome trajectories across trials (Figure 2b). The actual outcome on each trial was randomly sampled from probability distributions based on the generated means. Participants were informed about the volatile nature of the task environment, specifically regarding the fluctuation of both mean and the sampled rewards (see Appendix B for task instructions).

STAI, ECG, and Saliva Measures

STAI, ECG, and saliva measures were conducted but will not be discussed in the current thesis. ECG activity and salivary α-amylase levels served as manipulation check and correlates of central NA release.

EEG

Data Collection. EEG data was collected during the bandit task with a sampling rate of 512 Hz using a 32-channel electrode cap and the ActiveTwo System (BioSemi, Netherlands). The electrode cap layout followed the international 10/20 system, with an additional channel at the FCz position instead of T8. The activity of 32 scalp electrodes and of six exterior electrodes were recorded. The four electrodes surrounding the eyes served to measure eye movement, with electrodes below and above the eye capturing vertical eye movement (i.e., eye blinks) and electrodes at the outer canthi of both eyes detecting horizonal eye movement (i.e., saccades). The two exterior electrodes at the mastoid sites served as offline reference point.

Pre-processing. The EEG data was pre-processed in MATLAB (Version 2021a) using the EEGLAB toolbox (Delorme & Makeig, 2004). First, a bandpass filter to include frequencies between 0.1 Hz and 35 Hz using the EEGLAB filter method firls was applied and the data was re-referenced to the mastoids. Then, we divided the data into epochs, starting 2000 ms before and ending 2500 ms after feedback onset and removed baseline activity starting 200 ms before feedback onset. Next, electrodes were interpolated if any of the three criteria applied: the kurtosis (5 *SD* cut-off), the joint probability (5 *SD* cut-off), and the spectrum criterion (3 *SD* cut-off). However, electrodes typically contaminated with eye blinks (Fp1, Fp2) were not interpolated since eye blink activity was corrected for separately. In total, 99 electrodes (5.5%) were interpolated, on average 1.53 electrodes (*SD* = 0.94) in the atomoxetine and 1.77 (*SD* = 1.38) in the placebo condition. The number of interpolated electrodes ranged from zero to five per participant and session, thus there was no excessive interpolation of electrodes (cut-off criterion = 20%, i.e., six electrodes per EEG recording session). As a next step, epoch segments were automatically rejected if the amplitude exceeded 300 µV or if the joint probability criterion (5 *SD* cut-off) applied. Both criteria were applied to the signal of the scalp electrodes (excluding those that are typically contaminated by eye blinks: Fp1, Fp2, AF3, and AF4) with a time window of 1000 ms before and 1500 ms after feedback onset. On average, 39.57 trials (range: 12-155) were rejected per participant in the atomoxetine condition, and 24.23 trials (range: 13-44) in the placebo condition. None of the EEG recordings had to be excluded, since the amount of rejected trials per participant and session did not exceed 25% (= 168 trials). As a result, clean data of 632.43 trials (*SD* = 34.27) per participant remained on average in the atomoxetine and 647.77 trials (*SD* = 8.39) in the placebo condition. Lastly, we run an independent component analysis (ICA; Bell &

Sejnowski, 1995) with the data of all scalp electrodes and rejected components that contained less than 5% of brain activity but more than 80% of artifact activity in total (categorised as muscle, eye, heart, line, or channel noise activity) using the EEGLAB plugin ICLabel.

Specifying the Spatial and Temporal Occurrence of Valence-Sensitive ERP Components. Our task design differed from bandit tasks typically used in the reward learning literature in that it employed continuous (1-99 points) instead of binary feedback (e.g., win or loss), and it entailed an additional complete feedback condition next to the more common partial feedback condition. For this reason, we decided to use a data-driven instead of a literature-driven approach to identify the spatial and temporal occurrence of the FRN and P3 or their approximation. Given the well-established sensitivity of the FRN and P3 to reward valence (Kirsch et al., 2022), we employed a difference wave approach to identify relevant time windows and electrode locations. Ultimately, our goal was to specify the time windows and locations in the ERP data for our later regression analyses (see section "Linking the ERP and Computational Data" of the Methods section).

To determine the spatial and temporal characteristics of the ERP components differentially sensitive to reward valence, we analysed the ERP signal by computing the difference between low reward trials (≤ 50 points, representing negative valence) and high reward trials (> 50 points, representing positive valence). We generated topography plots for each 50 ms interval, starting from feedback onset and extending to 1000 ms post-feedback. The plots revealed a pronounced negativity at around 400 ms and a strong positivity at around 700 ms both localized at parietal electrode sites (Figure 3a). Based on these results, we plotted the averaged ERP waveforms of the parietal cluster (including electrode Pz and neighbouring electrodes CP1, CP2, P3, P4, PO3, and PO4) to examine the precise temporal course of the parietal ERP signal and determine the time windows for the regression analyses (Figure 3b). Our analysis of the difference wave identified an early parietal negative deflection occurring between 350 and 450 ms and a late parietal positive deflection occurring between 600 and 800 ms. We refer to these early and late parietal components of the difference wave as the valence N400 and valence P700, respectively. Importantly, the characteristics of these identified components align with those of the FRN and P3, as the FRN is an early component exhibiting a strong negative difference wave deflection, while the P3 is a late component showing a positive difference wave deflection. For the main ERP analyses, however, we will not use the difference wave, but the specified time windows (350-450 ms and 600-800 ms) and electrodes (Pz, CP1, CP2, P3, P4, PO3, and PO4) to further determine the dependent variables reflecting "early parietal activity" (between 350 and 400 ms) and "late parietal activity" (between 600 and 800 ms). We will refer to these parietal cluster activities as Pz400 and Pz700, respectively, where Pz refers to the central cluster electrode, and 400 and 700 refer to the mean latency of the previously identified maximum activity after feedback onset.

Data Analyses

Data Exclusion Criteria

Participants that performed below chance level (i.e., indication that they did not learn) were excluded from all analyses. Performance accuracy was defined based on the previously generated mean reward trajectories; if participants chose the option that corresponded to the highest mean reward, their decision was classified as correct. For every block, the proportion of accurate decisions was calculated to determine performance. Given the original findings of the mean performance (*M* = 65%, *SD* = 5%), chance level was set to *M*-2**SD* = 55% for our study. We did not have to exclude any participants, since the participants' objective performance was above chance level (*M* = 66.36%, *SEM* = 0.61%, range = 58.26-72.17%, *t*-test against 50%: *t*(29) = 26.6749, *p* < .001).

Computational Analyses

Computational Modelling. To mathematically describe the action value learning process, we used the identical reinforcement learning models as Findling et al. (2019) did: the Q-learning model based on the original Rescorla-Wagner equation (further called "exact RL model"), and the noisy RL model as introduced by Findling et al. (2019). The noisy RL model differs from the exact RL model in entailing an additive random noise variable ε_t as can be seen in Equation 1.

$$Q_t = Q_{t-1} + \alpha(r_t - Q_{t-1}) + \varepsilon_t \tag{1}$$

Both RL models describe the action value (Q_t) updating process on each trial *t*, where α is the learning rate and r_t is the obtained reward. In these models lies the idea that the update of the action value (i.e., learning) is only necessary when the prediction error (PE = r_t - Q_{t-1}) does not equal zero, meaning that the reward was unexpected in one way or the other and differs from the previously estimated action value Q_{t-1}. Crucially, while the exact RL model assumes that the learning process is noise-free ($\varepsilon_t = 0$), the noisy RL model does not ($\varepsilon_t > 0$). The noisy RL model adds random noise to the equation to corrupt the action values, attempting to model not exact but imprecise learning. The noise variable ε_t is drawn from a normal distribution with zero mean and a standard deviation σ_t defined by a constant learning imprecision parameter ζ multiplied by the magnitude of the prediction error as described in Equation 2. Is the learning imprecision parameter set to zero, the random noise added to the equation will be zero, resulting in the exact RL model. Thus, Equation 1 describes both the exact and the noisy RL model. Unmentioned so far, to model the forgetting process we also included a learning decay parameter (ranging between 0 and 1) that was applied to the Q-values that were not being updated at the given trial. However, this is outside the scope of the thesis and will not be discussed further.

$$\sigma_t = \zeta|r_t - Q_{t-1}| \quad (2)$$

To model the action selection process capturing explorative and exploitative decision making, we used a stochastic softmax policy as described in Equation 3.

$$P(A_t) = \frac{e^{\frac{1}{\tau}Q_{At}}}{e^{\frac{1}{\tau}Q_{At}} + e^{\frac{1}{\tau}Q_{Bt}}} \quad (3)$$

The softmax rule converts the action values (Q_A for option *A* and Q_B for option *B*) into action probabilities, here the probability to choose option *A* at trial *t*: $P(A_t)$. The parameter τ describes the temperature and determines how strongly the choices are guided by the action values. The higher the τ the more stochastic the choices will be, reflecting explorative behaviour where choices are more likely to deviate from the option with the highest estimated Q-value (non-greedy decisions). In contrast to that, the lower the τ the more deterministic the choices will be, reflecting exploitative behaviour where the option with the highest estimated Q-value is always chosen to maximise short-term pay-off (greedy decisions).

Taking these two equations together, we created three distinct models to be able to test our hypothesis concerning the origin of non-greedy decisions. The first model captured exploration as the only source (exact RL model with softmax policy, $\tau > 0$, $\zeta = 0$), the second model captured both exploration and learning imprecision (noisy RL model with softmax policy, $\tau > 0$, $\zeta > 0$), and the third model captured only learning imprecision as a source for non-greedy decisions (noisy RL model with argmax policy, $\tau = 0$, $\zeta > 0$). The performance of these three models will be tested against each other. For clarification, the model parameters (including learning rate, learning decay, learning imprecision, and choice stochasticity) remain constant within each single condition, participant, and RL model, while the model variables (including prediction error, learning noise, and Q-value) vary across trials.

Model Fitting, Selection, and Simulation. To estimate the parameter set (learning rate, learning imprecision, and choice stochasticity) for every participant, condition, and RL model (exact RL-softmax, noisy RL-softmax, and noisy RL-argmax) separately, we used a Bayesian optimization algorithm, the Bayesian adaptive direct search method (BADS; Acerbi & Ma, 2017) which is based on variational Bayesian Monte Carlo sampling. We set the parameter bounds to the following values: learning rate α = [0.001 0.999], learning decay δ = [0 0.999], learning imprecision ζ = [0 10], and choice stochasticity τ = [10^{-12} 1].

In a next step, we were interested to know which of the models fitted the data best (H1a) by using a Bayesian model selection (BMS) procedure as implemented in the spm_bms function of the SPM12 toolbox (https://github.com/spm/spm12) with the previously estimated log likelihoods as input. This resulted in two model comparison metrics - the posterior probability $P_{\text{posterior}}$ and the exceedance probability $P_{\text{exceedance}}$. The posterior probability indicates how likely the model is given the data, while the exceedance probability indicates how more likely the model is than any of the other models given the data (Stephan et al., 2009). Thus, the higher the probabilities, the better the model fit compared to the alternative models.

Based on these two metrics, we chose the best fitting model: the noisy RL model with softmax policy, entailing a learning noise and exploration parameter. The estimated learning imprecision and choice stochasticity parameters of that model were later used as independent variables to test if increased NA levels are linked to increased learning imprecision but not to increased choice stochasticity (H2; see next section “Mixed-Effects ANOVAs of the Computational Data”). Moreover, we aimed to quantify the proportion of non-greedy decisions attributed to imprecise learning rather than to the need for exploration (H1b). To achieve this, we initially identified non-greedy trials based on the exact RL model. Subsequently, we examined whether any of these trials would be categorised as greedy by the noisy RL model with softmax policy. If such trials were identified, they would contribute to the fraction of non-greedy decisions accounted for by learning imprecision.

Lastly, we used the best fitting model (noisy RL model + softmax) for the model simulation process to recover the model variables such as the estimated action values (Q-values), the prediction error, and learning noise on the trial level. The simulated model variables were later used as regressor variables to test our third hypothesis H3b (see section "Exploratory Model-Based Analyses of the ERP Components").

Mixed-Effects ANOVAs of the Computational Data

To test if increased NA levels are linked to increased learning imprecision but not to increased choice stochasticity (H2), we ran two separate mixed-effects ANOVAs with treatment condition (main interest) and feedback condition (control) as within-subject factors, treatment order (control) as between-subject factor, and the respective model parameter (learning imprecision and choice stochasticity) as dependent variable. Treatment order was included to control for its potential effect but will not be reported or discussed further.

Linking the ERP and Computational Data

Exploratory Model-Free Analyses of the ERP Components. To explore if some model-free variables (i.e., variables that are not based on the RL models) are linked to neural activity (H3a), we ran two separate repeated measures ANOVAs. Feedback type (partial vs. complete) and reward magnitude (low: ≤ 50 points vs. high: > 50 points) served as our model-free independent variables. To operationalise neural activity serving as dependent variable for the ANOVAs, we employed a grand-average ERP approach to our EEG data as discussed earlier (for details, see section "Specifying the Spatial and Temporal Occurrence of Valence-Sensitive ERP Components"). In general, the dependent variable reflected average Pz400 and average Pz700 activity, respectively. More specifically, for every participant, we split the EEG data among the four conditions (partial – low, partial – high, complete – low, complete – high) to be able to test if feedback type and reward magnitude differentially affect the neural signal. Then, we averaged the EEG signal (per component, participant, and condition) across all cluster electrodes (Pz, CP1, CP2, P3, P4, PO3, and PO4), across the time points of the respective time window (Pz400: 350 to 450 ms; Pz700: 600 to 800 ms), and across all the trials (6 blocks à 56 trials per feedback condition, without any previously excluded trials) resulting in one amplitude value per ERP component, participant, and model-free condition (feedback x reward magnitude).

Exploratory Model-Based Analyses of the ERP Components. To explore if some of our central RL model variables predict neural activity (H3b), we conducted multi-level analyses for each component separately. The first level consisted of single-trial regression analyses examining the relationship between model variables and neural activity on a single-trial level, while the second level involved mixed-effects ANOVA analyses investigating the effect of feedback and treatment conditions on the standardised regression coefficients. As in the model-free analyses, we operationalised neural activity by employing a grand-average ERP approach to our EEG data (for details, see section “Specifying the Spatial and Temporal Occurrence of Valence-Sensitive ERP Components”). In general, the dependent variable reflected average parietal cluster activity in the time window of valence N400 (i.e., Pz400 amplitude) and valence P700 (i.e., Pz700 amplitude) on a single trial. To obtain the single-trial amplitudes for the regression analyses, we averaged the EEG signal for every participant across the parietal electrodes Pz, CP1, CP2, P3, P4, PO3, and PO4 and across the data timepoints from 350 to 450 ms (for activity within the valence N400 time window) and from 600 to 800 ms (for activity within the valence P700 time window) for each feedback and treatment condition separately. We specified the following model variables on the single-trial level as regressor variables: the signed reward prediction error, choice difficulty, and learning noise. Prediction error and learning noise were derived from the model simulation procedure directly (see section "Model Fitting, Selection, and Simulation"). Choice difficulty was defined as the absolute value of the difference of the simulated Q-values for the two options, where a smaller value reflects a more difficult choice that has to be made. The resulting linear equations for each participant and for each condition separately looked like the equation below (4), where each variable reflects values on a single-trial level:

$$EEG_{Pz400/Pz700} \sim \beta_0 + \beta_1 \times RPE + \beta_2 \times DIFFICULTY + \beta_3 \times NOISE + interaction\ terms + \varepsilon \tag{4}$$

This resulted in four single-trial regression models (one per condition: treatment x feedback) per ERP component and participant. To test if any of the regressor variables significantly predict activity in the time windows of valence N400 or valence P700, we tested the values of the three main regression coefficients (without the interaction terms) - averaged across conditions - against zero using the two-tailed *t*-test with a significance level α = .05 for each regression coefficient separately.

In a second step, we wanted to know if any of the effects are modulated by the NA levels or by how feedback is presented (H3c). Therefore, we ran a mixed-effects ANOVA for each regressor (RPE, choice difficulty, and learning noise) separately, with the respective regression coefficient variable as dependent variable. Treatment and feedback condition served as within-factor, whereas treatment order was included as a control as between-factor variable and will not be discussed further.

Results

No participants were excluded from the analyses, since they performed above chance level in the restless bandit task, indicating that they did learn (for details, see section "Data Exclusion Criteria" of the Methods section).

Performance and Comparison of the Reinforcement Learning Models

First, we ran the Bayesian model fitting procedure using variational Bayesian Monte Carlo sampling to estimate the model parameter set (Table 1). To test which of the three RL models performed best in each condition, we applied the BMS procedure with the log likelihood estimates as input. All relevant model comparison metrics are listed in Table 2. We used the posterior and the exceedance probability to demonstrate our findings (for explanation of the metrics, see section “Model Fitting, Selection, and Simulation” of the Methods section). As expected, the model entailing a learning imprecision and exploration (i.e., choice stochasticity) parameter explained the data best (Figure 4) in the partial feedback conditions of both treatment conditions (atomoxetine: $P_{\text{exceedance}} = 1$; placebo: $P_{\text{exceedance}} = 1$), but unexpectedly also in the complete feedback conditions (atomoxetine: $P_{\text{exceedance}} > 0.99$; placebo: $P_{\text{exceedance}} = 1$). Given these results, we used the noisy RL model with softmax action selection rule for further analyses.

Next, we were interested in the fraction of non-greedy decisions that can be explained by learning imprecision. Using the same approach as Findling et al. (2019), our analysis yielded consistent results. As expected, participants made a higher fraction of non-greedy decisions due to learning imprecision in the complete feedback conditions (atomoxetine: 75.72%; placebo: 82.15%) than in the partial feedback conditions (atomoxetine: 51.88% ; placebo: 45.04%). These results imply that the fraction of non-greedy decisions due to choice stochasticity is lower in the complete than in the partial feedback condition, showing that less exploration is exerted when both outcomes are revealed.

Effect of Noradrenaline on Learning Imprecision and Choice Stochasticity

To test if increased NA levels are linked to changes in learning imprecision and choice stochasticity (i.e., exploration), we ran two separate mixed-effects ANOVAs. We were interested in the two model parameters since learning imprecision controls for the amount of learning noise generated on each trial, while choice stochasticity determines how strongly the decisions are guided by the estimated action values. Contrary to our hypothesis that increased NA levels are linked to increased learning imprecision but not to increased choice stochasticity, the results of the analyses (Table 3) did not reveal a main effect of the treatment condition on learning imprecision, $F(1, 29) = 0.37$, $p = .549$, meaning that increased NA was not linked to increased learning imprecision. However, as expected, increased NA levels did not impact choice stochasticity either, $F(1, 29) = 0.15$, $p = .704$. Interestingly, there was a main effect of feedback type on choice stochasticity, $F(1, 29) = 19.39$, $p < .001$. Participants explored on average more when only the outcome of their chosen option was presented to them (atomoxetine: $M_\beta = 0.06$, $SD = 0.01$; placebo: $M_\beta = 0.07$, $SD = 0.01$) than when both outcomes were revealed (atomoxetine: $M_\beta = 0.03$, $SD = 0.01$; placebo: $M_\beta = 0.02$, $SD = 0.01$). This finding is consistent with the notion of our first hypothesis, since it suggests that exploration is lower, albeit not inexistent, in the partial outcome condition than in the complete outcome condition.

Exploratory Analyses of the ERP Data

Effect of Feedback Type and Reward Magnitude on Pz400 and Pz700 Amplitude

First, we wished to explore if some model-free variables (i.e., variables that are not based on the RL models) are linked to neural activity. To operationalise neural activity, we used a grand-average approach (for details, see section “Exploratory Model-Free Analyses of the ERP Components” of the Methods section), resulting in two main cluster activities at parietal site with peak activity between 350 and 450 ms (Pz400) and between 600 and 800 ms (Pz700) after feedback onset (Figure 3). Since we wanted to investigate if feedback type and reward magnitude influence Pz400 and Pz700 amplitude, we ran two separate repeated measures ANOVAs with feedback type (partial vs. complete) and reward magnitude (low: ≤ 50 points vs. high: > 50 points) as independent variables.

The model-free analysis for the early parietal component (Pz400; Figure 5) revealed a significant large main effect of feedback type on the mean parietal amplitude between 350 and 450 ms after feedback presentation, *F*(1, 29) = 10.02, *p* = .002, partial η^2 = 0.14, but no significant main effect of reward magnitude, *F*(1,59) = 1.79, *p* = .186, partial η^2 = 0.03, or interaction effect of the two, *F*(1,29) = 2.20, *p* = .144, partial η^2 = 0.04. The average Pz400 amplitude was significantly higher when feedback was presented partially (*M* = 6.83 mV, *SD* = 5.29) compared to when feedback was presented completely (*M* = 5.03 mV, *SD* = 4.87).

The model-free analysis for the late parietal component (Pz700; Figure 6) revealed a large main effect of reward magnitude on the mean parietal amplitude between 600 and 800 ms after feedback presentation, *F*(1,29) = 19.01, *p* < .001, partial η^2 = 0.24. There was no significant main effect of feedback type, *F*(1,29) = 0.42, *p* = .521, partial η^2 = 0.01, nor of the interaction of the two, *F*(1,29) = 0.12, *p* = .725, partial η^2 = 0.00. The average Pz700 amplitude was significantly higher when reward was low (reward = 1-50 points; *M* = 7.27 mV, *SD* = 6.08) compared to when reward was high (reward = 51-99 points; *M* = 6.41 mV, *SD* = 6.03).

Effect of Prediction Error, Choice Difficulty, and Learning Noise on Pz400 and Pz700 Amplitude and the Modulation thereof by Noradrenaline and Feedback Type

Next, we wanted to know if some of our central RL model variables predict single-trial neural activity by applying regression analyses for each condition (treatment x feedback) and participant separately. The dependent variable reflecting single-trial neural activity was operationalised using a grand-average approach (for details, see section "Exploratory Model-Based Analyses of the ERP Components"), resulting in two main cluster activities at parietal site with peak activity between 350 and 450 ms (Pz400) and between 600 and 800 ms (Pz700) after feedback onset (Figure 3). The central model variables serving as regressor variables were reward prediction error (signed), learning noise, and choice difficulty (|(Q_A – $Q_{B|}$) on the single-trial level. After obtaining the regressor coefficients for each participant and condition separately, we tested the mean value of the standardised regression coefficients (averaged across conditions) against zero to check if the regressors significantly predict parietal activity within the time window of valence N400 (i.e., Pz400) and valence P700 (i.e., Pz700). The findings show that all model variables significantly predict Pz400 and Pz700 amplitude on a single-trial level (Table 4), however all to a different extent. Choice difficulty predicts Pz400 and Pz700 activity the strongest, followed by reward prediction error, and learning noise.

Subsequently, we were interested to test if any of the effects is modulated by NA levels or feedback type for which we ran a mixed-effects ANOVA (with treatment and feedback condition as within- and treatment order as between-factor). The standardised regression coefficients across the four conditions (treatment x feedback) are illustrated in Figure 7 for Pz400, and in Figure 8 for Pz700.

The mixed-effects ANOVA for the early component (parietal activity within the time window of valence N400) yielded significant effects of feedback type on the regression coefficient of the reward prediction error and learning noise, but not of choice difficulty (Table 5). The effect of reward prediction error on Pz400 amplitude was stronger when feedback was presented partially compared to completely, main effect of feedback type on $\beta_{\mathrm{N400{\sim}RPE}}$: $F(1,28) = 5.37$, $p = .028$. The effect of learning noise on Pz400 amplitude on the other hand was stronger when both feedback outcomes were revealed, main effect of feedback type on $\beta_{\mathrm{N400{\sim}NOISE}}$: $F(1,28) = 5.39$, $p = .028$). The effect of choice difficulty on Pz400 amplitude, $F(1,28) = 0.05$, $p = .827$, was not modulated by the feedback condition, and the treatment condition did not have any significant effect on the regression coefficients.

The mixed-effects ANOVA for the late component (parietal activity within the time window of valence P700) yielded significant effects of feedback type on the regression coefficients of choice difficulty and learning noise, but not of prediction error (Table 6). The effect of choice difficulty on Pz700 amplitude was stronger when feedback was presented partially compared to completely, main effect of feedback type on $\beta_{\text{P700\textasciitilde DIFFICULTY}}$: $F(1,28) = 5.65$, $p = .024$. The effect of learning noise on Pz700 amplitude on the other hand was stronger when both feedback outcomes were revealed, main effect of feedback type on $\beta_{\text{P700\textasciitilde NOISE}}$: $F(1,28) = 9.85$, $p = .004$. The effect of reward prediction error on Pz700 amplitude was not modulated by the feedback condition, $F(1,28) = 2.62$, $p = .117$, and the treatment condition did not have any significant effect on the regression coefficients.

Discussion

The main aim of the current study was to investigate the role of the LC-NA system in modulating learning imprecision during value-based learning by using a multi-modal approach including reinforcement learning, a pharmacological manipulation, and EEG data. The computational results indicate that the RL model entailing a learning imprecision and exploration (i.e., choice stochasticity) parameter explained the human learning and decision-making data better than the models with only or without imprecision parameter. However, contrary to our expectations, this was also the case when counterfactual information was available, and exploration was hypothesised not to be needed. On the contrary, we did find a higher fraction of non-greedy decisions due to learning imprecision in the complete compared to the partial feedback condition, suggesting that less explorative decisions were made when counterfactual information was presented. This was confirmed by further analyses, in which we found a main effect of feedback type on choice stochasticity, showing that participants explored less in the complete compared to the partial feedback condition. However, our main expectation of the second hypothesis could not be confirmed since no evidence was found for an association between learning imprecision and NA levels. Lastly, our exploratory approach of the ERP data revealed that the amount of presented information affected the early parietal cluster activity between 350 and 450 ms after feedback presentation (Pz400), while reward magnitude affected the later parietal cluster activity between 600 and 800 ms after feedback presentation (Pz700). Moreover, both parietal components were significantly predicted by choice difficulty, prediction error, and learning noise. NA did not modulate any of these effects, while the amount of presented feedback did modulate the effects of the prediction error on Pz400, choice stochasticity on Pz700, and learning noise on both components.

Our main computational findings suggest that learning imprecision plays a crucial role in human value-based learning and decision-making. This challenges the idea of the origin of non-greedy decisions that fail to maximise short-term reward. Previous literature has attributed non-greedy decisions to exploration only (Aston-Jones & Cohen, 2005; Daw et al., 2006; Dayan & Daw, 2008), suggesting that behavioural variability arises solely from the exploration-exploitation dilemma. Importantly, our findings demonstrate that learning imprecision is another source of non-maximising decisions, in addition to the need to explore. This partially aligns with prior findings of the study we wished to replicate (Findling et al., 2019). In accordance with the original findings, the fraction of non-greedy decisions explained by learning imprecision was higher when only information on the chosen outcomes was revealed. However, contrary to the original findings, exploration was not fully eliminated when counterfactual information was available. Thus, the manipulation of the amount of feedback (complete vs. partial) did not result in eliminating the need for exploration during value-based decision tasks. Subsequently, the results of Findling et al. (2019) showed that the noisy RL model with learning imprecision but not with choice stochasticity explained the data best in the complete feedback conditions, while our results suggest that the noisy RL model with both parameters explained the data best, irrespective of the feedback amount. Despite the discrepancies to the findings of Findling et al. (2019), the decomposition of non-greedy decisions in our study did reveal a potential trend towards the original findings since exploration was lower during complete compared to partial feedback conditions. Alternatively, a possible explanation for our contradictory findings may lie in the volatile nature of the decision environment. Previous research has shown that volatility drives exploration (Jepma et al., 2020). Accordingly, the uncertainty inherent in such volatile environments may increase the importance of not relying solely on estimations of the highest reward. Thus, choice stochasticity continues to play a role even when counterfactual information is available, particularly in scenarios where the reward-outcome contingencies undergo rapid shifts or when the sampled rewards strongly deviate from the underlying reward structures, as observed in our restless bandit task.

Contrary to our expectations, higher NA levels did not affect learning imprecision, however as expected were also not linked to choice stochasticity. This is further evidence that the LC-NA system is not involved in balancing exploration against exploitation as suggested by the adaptive gain theory (Aston-Jones & Cohen, 2005). Already earlier studies could not find conclusive evidence for the link between high tonic NA levels and exploration, showing either no association (Jepma et al., 2010) or an association in the opposite direction as suggested by the adaptive gain theory (i.e., higher tonic NA levels were correlated with less exploration; Warren et al., 2017). It is important to note, however, that these three studies have notable differences making the comparison of the findings difficult. While Warren et al. (2017) used the same dose and substance (40mg of atomoxetine) to increase tonic NA levels in a within-subject design, the timing of the administration (180 minutes before the decision task), the decision task (a modified version of the horizon task as described by Wilson et al., 2014), and the decision environment (stable reward-outcome contingencies) was different from our study. On the other hand, Jepma et al. (2010) used a similar decision task (4-armed bandit task) with volatile reward-outcome contingencies as well, but with a between-subject study design and the administration of 4mg of reboxetine 2 to 3 hours prior the gambling task to increase tonic NA levels. Hence, a closer examination of the biochemical aspects of the substances is required. This point will be addressed in more detail later.

In terms of ERP correlates of value-based learning and decision-making, we identified two valence-sensitive ERP components using a data-driven difference wave approach to identify approximates of the FRN and P3. We identified the highest activity of the difference wave between positive minus negative reward to be at parietal sites (electrode Pz and neighbouring electrodes) with a peak negative deflection between 350 and 450 ms (valence N400) and a peak positive deflection between 600 to 800 ms (valence P700), respectively. Most literature identifies peak activity of the FRN and P3 difference waves to occur earlier, namely 200 to 300 ms and 300 to 600 ms after feedback presentation, respectively (for review, see Kirsch et al., 2022). Moreover, peak activity of the FRN has been identified frontocentrally at the FCz electrode, while we found maximum activity of the valence N400 more posterior at parietal sites (electrode Pz and neighbouring electrodes). Nevertheless, spatial and temporal locations of the FRN and P3 are variable, and other research found similar results as we did (e.g., Balconi & Crivelli, 2010). Thus, averaged EEG activity of the parietal clusters within the respective time windows made a good approximation of the FRN and P3, further used in the main ERP analyses.

In accordance with the two-step theory of feedback evaluation (Bernat et al., 2015), our results show an influence of reward magnitude on the late parietal cluster activity within the time window of valence P700 (i.e., Pz700 amplitude) but not the early parietal cluster activity within the time window of valence N400 (i.e., Pz400 amplitude), suggesting that a more in-depth evaluation of the feedback (here the evaluation of reward magnitude) happens later. Given that we used reward valence (positive: ≤ 50 points vs. negative: > 50 points) in our ERP approach to identify the temporal and spatial sites of valence-sensitive ERP components (i.e., valence N400 and valence P700), the null finding of reward magnitude on the Pz400 amplitude is counterintuitive. A possible explanation may lie in the fact, that the difference wave of reward valence is less pronounced for valence N400 than for valence P700, as can be seen in the topography and the ERP plot of the difference wave. Concerning the effect of feedback type on Pz400 and Pz700 amplitude, the amount of information revealed (partial vs. complete) had only an influence on the Pz400 amplitude but not on Pz700 amplitude. These findings may suggest that processing the type of feedback delivery (partial vs. complete) suffices superficial elaboration during the early stage. Additionally, when information is missing, the Pz400 amplitude tends to be higher compared to when all information is revealed. This may imply an increased need to process the uncertainty about the reward-outcome contingencies that arises from missing information.

Interestingly, we found that the signed reward prediction error (RPE), choice difficulty, and learning noise significantly predicted both parietal cluster activities (within the time window of valence N400 and valence P700) on a single-trial basis. These findings are not conforming previous literature suggesting that the FRN scales with a signed RPE (Holroyd & Coles, 2002; Kirsch et al., 2022), while the P3 scales with an unsigned RPE (Mars et al., 2008). Our main interest lied in investigating the underlying neural dynamics of learning noise, however. Even though we did find that single-trial learning noise significantly predicted averaged single-trial EEG activity between 350 and 450 ms and between 600 and 800 ms at parietal sites, choice difficulty predicted both cluster activities the strongest, followed by the reward prediction error. Moreover, the strength of prediction of learning noise was the same for the early and late component. These findings may imply that learning imprecision manifests in a broader, more generalised manner, as suggested by Findling & Wyart (2021), rather than during a single process within a certain time window. The authors argue that genuine neural noise could arise from stochastic synaptic release at the cellular level, from random fluctuations in the delicate balance between excitation and inhibition that neural populations need for precise computations, or the variable process of pooling task-relevant information.

Contrary to our primary hypothesis involving the role of the LC-NA system in modulating learning imprecision, we were unable to establish a connection between elevated NA levels and learning imprecision. Furthermore, the relationship between learning imprecision and neural activity was not moderated by central NA release. These null findings may raise questions about the pharmacological manipulation used in the current study, posing a potential limitation and implications for future research on that topic. For instance, evidence on the pharmacokinetics of atomoxetine (Sauer et al., 2005) shows individual differences in metabolising atomoxetine. Accordingly, "slow" metabolisers demonstrate an approximately 10-fold higher average steady-state plasma concentration of atomoxetine than "extensive" metaboliser, implicating significant differences in how and when the effect of atomoxetine unfolds. Additionally, baseline dependent effects may contribute to different treatment effects depending on the individual baseline NA activity, similar to the implications of the inverted-U relationship between tonic LC activity and task performance suggested by Aston-Jones and Cohen (2005) or the findings on tonic LC-NA activity and trait anxiety (Howells et al., 2012). Lastly, the NA system is intricately interconnected with other neurotransmitter systems, such as dopamine (Ranjbar-Slamloo & Fazlali, 2020). For instance, the NA transporter has been found to be responsible for dopamine reuptake in the cortex as well (Devoto & Flore, 2006). Thus, a single dose of atomoxetine may not only increase central NA levels, but also cortical dopamine levels. This, and the fact that neurotransmitter systems often have feedback mechanisms that regulate their activity to maintain balance may mitigate the effect of central NA release due to atomoxetine. Interestingly, a recent re-analysis of the study by Jepma et al. (2010), which initially demonstrated no association between exploration and central NA levels following the administration of 4 mg of reboxetine (another NA reuptake inhibitor), has yielded positive findings regarding the link between reboxetine and computational learning imprecision (Findling & Wyart, 2021). The reboxetine group exhibited a larger fraction of behavioural variability (80%) compared to the placebo group (53%). Furthermore, computational learning noise estimates were higher in the reboxetine group compared to the placebo group, indicating an increase in learning imprecision due to the administration of a low dose of reboxetine. Thus, despite the present null findings regarding elevated NA levels and learning imprecision, future research should consider using reboxetine to increase central NA levels as an alternative to atomoxetine and assess potential individual differences and baseline dependent effects of the pharmacological substance.

Taken together, the current study did not find evidence for a modulatory role of the noradrenergic locus coeruleus system on learning imprecision or choice stochasticity during human value-based learning and decision-making in volatile environments. Nevertheless, the present research provides further evidence for the importance of learning imprecision in human value-based learning. Learning imprecision contributes to a substantial number of non-greedy decisions failing to maximise reward, especially when counterfactual information is not available. The noisy reinforcement learning model incorporating a learning imprecision parameter outperforms the classical reinforcement learning model that assumes exact learning. On a neural level, our exploratory data suggests an early and a late, more in-depth, feedback evaluation process, further implying that computational noise emerges more generally during feedback processing rather than during a certain time of feedback processing. Future research using reboxetine and accounting for individual differences in metabolising noradrenaline may help bridge the gap between learning imprecision and potential underlying neural mechanisms such as the LC-NA system.

References

Acerbi, L., & Ma, W. J. (2017). Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. *Advances in Neural Information Processing Systems*, *30*, 1834–1844. https://papers.nips.cc/paper/2017/hash/df0aab058ce179e4f7ab135ed4e641a9-Abstract.html

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*(1), 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Balconi, M., & Crivelli, D. (2010). FRN and P300 ERP effect modulation in response to feedback sensitivity: The contribution of punishment-reward system (BIS/BAS) and Behaviour Identification of action. *Neuroscience Research*, *66*(2), 162–172. https://doi.org/10.1016/j.neures.2009.10.011

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129

Beltzer, E. K., Fortunato, C. K., Guaderrama, M. M., Peckins, M. K., Garramone, B. M., & Granger, D. A. (2010). Salivary flow and alpha-amylase: Collection technique, duration, and oral fluid type. *Physiology & Behavior*, *101*(2), 289–296. https://doi.org/10.1016/j.physbeh.2010.05.016

Bernat, E. M., Nelson, L. D., & Baskin-Sommers, A. R. (2015). Time-frequency theta and delta measures index separable components of feedback processing in a gambling task. *Psychophysiology*, *52*(5), 626–637. https://doi.org/10.1111/psyp.12390

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. https://doi.org/10.1038/nature04766

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453. https://doi.org/10.3758/CABN.8.4.429

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Devoto, P., & Flore, G. (2006). On the origin of cortical dopamine: Is it a co-transmitter in noradrenergic neurons? *Current Neuropharmacology*, *4*(2), 115–125.

Drugowitsch, J., Wyart, V., Devauchelle, A.-D., & Koechlin, E. (2016). Computational precision of mental inference as critical source of human choice suboptimality. *Neuron*, *92*(6), 1398–1411. https://doi.org/10.1016/j.neuron.2016.11.005

Duncan-Johnson, C. C., & Donchin, E. (1977). On quantifying surprise: The variation of event-related potentials with subjective probability. *Psychophysiology*, *14*(5), 456–467. https://doi.org/10.1111/j.1469-8986.1977.tb01312.x

Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2019). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nature Neuroscience*, *22*(12), 2066–2077. https://doi.org/10.1038/s41593-019-0518-9

Findling, C., & Wyart, V. (2021). Computation noise in human learning and decision-making: Origin, impact, function. *Current Opinion in Behavioral Sciences*, *38*, 124–132. https://doi.org/10.1016/j.cobeha.2021.02.018

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112. https://doi.org/10.1007/BF02289823

Heil, S. H., Holmes, H. W., Bickel, W. K., Higgins, S. T., Badger, G. J., Laws, H. F., & Faries, D. E. (2002). Comparison of the subjective, physiological, and psychomotor effects of atomoxetine and methylphenidate in light drug users. *Drug and Alcohol Dependence*, *67*(2), 149–156. https://doi.org/10.1016/S0376-8716(02)00053-4

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. https://doi.org/10.1037/0033-295X.109.4.679

Howells, F. M., Stein, D. J., & Russell, V. A. (2012). Synergistic tonic and phasic activity of the locus coeruleus norepinephrine (LC-NE) arousal system is required for optimal attentional performance. *Metabolic Brain Disease*, *27*(3), 267–274. https://doi.org/10.1007/s11011-012-9287-9

Jepma, M., Schaaf, J. V., Visser, I., & Huizenga, H. M. (2020). Uncertainty-driven regulation of learning and exploration in adolescents: A computational account. *PLOS Computational Biology*, *16*(9), e1008276. https://doi.org/10.1371/journal.pcbi.1008276

Jepma, M., Te Beek, E. T., Wagenmakers, E.-J., Van Gerven, J. M. A., & Nieuwenhuis, S. (2010). The role of the noradrenergic system in the exploration-exploitation trade-off: A pharmacological study. *Frontiers in Human Neuroscience*, *4*, Article 170. https://doi.org/10.3389/fnhum.2010.00170

Kirsch, F., Kirschner, H., Fischer, A. G., Klein, T. A., & Ullsperger, M. (2022). Disentangling performance-monitoring signals encoded in feedback-related EEG dynamics. *NeuroImage*, *257*, Article 119322. https://doi.org/10.1016/j.neuroimage.2022.119322

Lindsay, G. (2021). *Models of the mind: How physics, engineering and mathematics have shaped our understanding of the brain*. Bloomsbury Sigma.

Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., & Bestmann, S. (2008). Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. *Journal of Neuroscience*, *28*(47), 12539–12545. https://doi.org/10.1523/JNEUROSCI.2925-08.2008

Nieuwenhuis, S. (2011). Learning, the P3, and the locus coeruleus-norepinephrine system. In R. Mars, J. Sallet, M. Rushworth, & N. Yeung (Eds.), *Neural basis of motivational and cognitive control* (pp. 209–222). Oxford University Press.

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus–norepinephrine system. *Psychological Bulletin*, *131*(4), 510. https://doi.org/10.1037/0033-2909.131.4.510

Polich, J. (2020). 50+ years of P300: Where are we now? *Psychophysiology*, *57*(7), Article e13616. https://doi.org/10.1111/psyp.13616

Rajkowski, J., Kubiak, P., & Aston-Jones, G. (1994). Locus coeruleus activity in monkey: Phasic and tonic changes are associated with altered vigilance. *Brain Research Bulletin*, *35*(5), 607–616. https://doi.org/10.1016/0361-9230(94)90175-9

Ranjbar-Slamloo, Y., & Fazlali, Z. (2020). Dopamine and noradrenaline in the brain; Overlapping or dissociate functions? *Frontiers in Molecular Neuroscience*, *12*. https://www.frontiersin.org/articles/10.3389/fnmol.2019.00334

Sambrook, T. D., & Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. *Psychological Bulletin*, *141*(1), 213–235. https://doi.org/10.1037/bul0000006

San Martín, R. (2012). Event-related potential studies of outcome processing and feedback-guided learning. *Frontiers in Human Neuroscience*, *6*, Article 304. https://doi.org/10.3389/fnhum.2012.00304

Sara, S. J. (2009). The locus coeruleus and noradrenergic modulation of cognition. *Nature Reviews Neuroscience*, *10*(3), Article 3. https://doi.org/10.1038/nrn2573

Sara, S. J., & Bouret, S. (2012). Orienting and reorienting: The locus coeruleus mediates cognition through arousal. *Neuron*, *76*(1), 130–141. https://doi.org/10.1016/j.neuron.2012.09.011

Sauer, J.-M., Ring, B. J., & Witcher, J. W. (2005). Clinical pharmacokinetics of atomoxetine. *Clinical Pharmacokinetics*, *44*(6), 571–590. https://doi.org/10.2165/00003088-200544060-00002

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

Segal, S. K., & Cahill, L. (2009). Endogenous noradrenergic activation and memory for emotional material in men and women. *Psychoneuroendocrinology*, *34*(9), 1263–1271. https://doi.org/10.1016/j.psyneuen.2009.04.020

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). Manual for the State-Trait Anxiety Inventory; Palo Alto, CA, Ed. *Palo Alto: Spielberger*.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). The MIT Press.

Thiele, A., & Bellgrove, M. A. (2018). Neuromodulation of attention. *Neuron*, *97*(4), 769–785. https://doi.org/10.1016/j.neuron.2018.01.008

Warren, C. M., Wilson, R. C., Wee, N. J. van der, Giltay, E. J., Noorden, M. S. van, Cohen, J. D., & Nieuwenhuis, S. (2017). The effect of atomoxetine on random and directed exploration in humans. *PLOS ONE*, *12*(4), Article e0176034. https://doi.org/10.1371/journal.pone.0176034

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074–2081. https://doi.org/10.1037/a0038199

Wyart, V., & Koechlin, E. (2016). Choice variability and suboptimality in uncertain environments. *Current Opinion in Behavioral Sciences*, *11*, 109–115. https://doi.org/10.1016/j.cobeha.2016.07.003

Yeung, N., & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *Journal of Neuroscience*, *24*(28), 6258–6264. https://doi.org/10.1523/JNEUROSCI.4537-03.2004

Tables

Table 1

Mean Values and Standard Deviations of the RL Model Parameters

	Model											
Condition	Exact RL softmax				Noisy RL argmax				Noisy RL softmax			
Treatment	Atomoxetine		Placebo		Atomoxetine		Placebo		Atomoxetine		Placebo	
Feedback	Partial	Complete	Partial	Complete	Partial	Complete	Partial	Complete	Partial	Complete	Partial	Complete
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)
Parameter												
Learning rate	.79	.39	.70	.44	.51	.50	.51	.50	.60	.54	.65	.54
	(.21)	(.19)	(.27)	(.25)	(.24)	(.22)	(.23)	(.19)	(.24)	(.20)	(.25)	(.21)
Learning	-	-	-	-	1.62	.97	2.27	1.17	.78	.69	.83	.88
imprecision					(1.17)	(.57)	(2.16)	(1.30)	(.98)	(.41)	(1.36)	(.89)
Choice	.10	.03	.12	.07	-	-	-	-	.06	.03	.07	.02
Stochasticity	(.08)	(.03)	(.18)	(.14)					(.07)	(.03)	(.06)	(.03)

Note. *N* = 30. The values were obtained after the model fitting procedure. The exact RL (reinforcement learning) softmax model contains learning rate and choice stochasticity (i.e., exploration) as model parameters, the noisy RL argmax model contains learning rate and learning imprecision, and the noisy RL softmax model contains all three model parameters (learning rate, learning imprecision, and choice stochasticity).

Table 2

Model Comparison Metrics of the RL Models

Condition	Model	*k*	*ll*	$P_{\text{posterior}}$	$P_{\text{exceedance}}$	AIC	BIC
Atomoxetine							
Partial	τ	3	-137.05	.04	0	280.20	291.56
			(37.92)			(75.84)	(75.84)
	ζ	3	-145.21	.05	0	296.49	307.87
			(41.82)			(83.64)	(83.64)
	ζ τ	4	-129.13	.91	1	266.38	281.53
			(37.17)			(74.34)	(74.34)
Complete	τ	3	-122.80	.09	0	251.67	263.05
			(37.00)			(73.99)	(73.99)
	ζ	3	-118.21	.18	< .01	242.49	253.87
			(38.49)			(76.99)	(76.99)
	ζ τ	4	-115.32	.73	> .99	238.77	253.91
			(35.61)			(71.22)	(71.22)
Placebo							
Partial	τ	3	-142.75	.04	0	291.57	302.95
			(44.48)			(88.95)	(88.95)
	ζ	3	-156.49	.03	0	319.05	330.43
			(46.73)			(93.45)	(93.45)
	ζ τ	4	-135.61	.93	1	279.35	294.50
			(42.84)			(85.67)	(85.67)
Complete	τ	3	-123.20	.09	0	252.46	263.84
			(43.69)			(87.38)	(87.38)
	ζ	3	-117.77	.13	0	241.61	252.99
			(43.95)			(87.91)	(87.91)
	ζ τ	4	-115.23	.78	1	238.58	253.72
			(43.34)			(86.68)	(86.68)

Note. Mean values and standard deviations of the log likelihood, AIC, and BIC values were derived across participants ($N = 30$). τ = model contains exploration parameter (exact RL softmax); ζ = model contains learning imprecision parameter (noisy RL argmax); ζ τ = model contains exploration and learning noise parameter (noisy RL softmax); k = number of model parameters; *ll* = log likelihood; $P_{\text{posterior}}$ = posterior probability; $P_{\text{exceedance}}$ = exceedance probability; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Table 3

Descriptive Statistics (Mean Values, Standard Errors) and the Mixed-Effects ANOVA Statistics for the Model Parameters Learning Imprecision and Choice Stochasticity (Temperature τ)

Model Parameter	Partial		Complete		ANOVA		
	M	*SE*	*M*	*SE*	Effect	*F*(1, 29)	*p*
Learning imprecision							
Atomoxetine	0.78	0.18	0.69	0.08	F	0.02	.882
Placebo	0.83	0.25	0.88	0.16	T	0.37	.549
					F x T	0.61	.441
Choice stochasticity							
Atomoxetine	0.06	0.01	0.03	0.01	F	19.39	< .001***
Placebo	0.07	0.01	0.02	0.01	T	0.15	.704
					F x T	1.12	.299

Note. *N* = 30. ANOVA = analysis of variance; Partial = partial feedback condition; Complete = complete feedback condition; F = feedback; T = treatment.

***p < .001.

Table 4

Regression Results of the Pz400 and Pz700 Amplitude for Reward Prediction Error, Choice Difficulty, and Learning Noise

Variable	β	*SE*	*t*(29)	*p*	95% CI
			N400		
RPE	0.67	0.12	5.80	< .001***	[0.43, 0.91]
Difficulty	-1.05	0.16	-6.77	< .001***	[-1.37, -0.73]
Noise	0.28	0.08	3.44	.002**	[0.11, 0.45]
			P700		
RPE	-0.33	0.11	-3.02	.005**	[-0.56, -0.11]
Difficulty	-1.56	0.20	-7.64	< .001***	[-1.98, -1.14]
Noise	0.28	0.12	2.43	.021*	[0.05, 0.52]

Note. For each participant (*N* = 30), we examined the impact of the single-trial RL model variables (RPE, choice difficulty, learning noise) on the averaged single-trial EEG activity within the time window of valence N400 (i.e., Pz400 amplitude) and the valence P700 (i.e., Pz700 amplitude) with separate regression analyses for the ERP components and conditions (feedback: partial vs. complete; treatment: atomoxetine vs. placebo). β = mean value of the standardised regression coefficients averaged across conditions and across participants; CI = confidence interval; RPE = reward prediction error. The results of the two-tailed *t*-tests indicate if the standardised regression coefficients averaged across the conditions (*N* = 30) are significantly different from zero (however, assuming non-normality). Both regressions (for Pz400 and Pz700) yielded the same mean value of the standardised regression coefficient of 0.28 for learning noise as regressor. See Table 5 (for Pz400) and Table 6 (for valence Pz700) for mean values and standard errors of the standardised regression coefficients across conditions.

*p < .05 **p < .01 ***p < .001.

Table 5

Descriptive Statistics (Mean Values, Standard Errors) and the Mixed-Effects ANOVA Statistics for the Standardised Regression Coefficients (of Reward Prediction Error, Choice Difficulty, and Learning Noise) for Parietal Activity Within the Time Window of Valence N400

	Partial		Complete		ANOVA		
Regression Coefficient	*M*	*SE*	*M*	*SE*	Effect	*F*(1, 29)	*p*
$\beta_{Pz400{\sim}RPE}$							
Atomoxetine	0.93	0.23	0.55	0.20	F	5.37	.028*
Placebo	0.84	0.17	0.36	0.16	T	0.64	.432
					F x T	0.10	.751
$\beta_{Pz400{\sim}DIFFICULTY}$							
Atomoxetine	-1.23	0.25	-1.14	0.26	F	0.05	.827
Placebo	-0.91	0.20	-0.93	0.19	T	1.73	.200
					F x T	0.08	.776
$\beta_{Pz400{\sim}NOISE}$							
Atomoxetine	0.08	0.17	0.51	0.16	F	5.39	.028*
Placebo	0.01	0.19	0.53	0.17	T	0.03	.867
					F x T	0.07	.796

Note. In a first step, we examined for each participant (*N* = 30) the impact of the RL model variables (RPE, difficulty, noise) on the single-trial parietal activity within the time window of valence N400 with separate regression analyses for the four conditions (feedback: partial vs. complete; treatment: atomoxetine vs. placebo). In a second step, we ran a mixed-effects ANOVA to examine the effect of feedback and treatment condition on the standardised regression coefficients (assuming equal variances across conditions, however, non-normality of the distributions). The Greenhouse-Geisser (1959) correction did not yield different *p* values. RPE = reward prediction error; ANOVA = analysis of variance; Partial = partial feedback condition; Complete = complete feedback condition; F = feedback; T = treatment. For illustration of the findings, see Figure 7.

$^*p < .05$

Table 6

Descriptive Statistics (Mean Values, Standard Errors) and the Mixed-Effects ANOVA Statistics for the Standardised Regression Coefficients (of Reward Prediction Error, Choice Difficulty, and Learning Noise) for Parietal Activity Within the Time Window of Valence P700

	Partial		Complete		ANOVA		
Regression Coefficient	*M*	*SE*	*M*	*SE*	Effect	*F*(1, 29)	*p*
$\beta_{\mathrm{Pz700{\sim}RPE}}$							
Atomoxetine	-0.20	0.23	-0.26	0.20	F	2.62	.117
Placebo	-0.16	0.19	-0.71	0.14	T	1.28	.267
					F x T	2.05	.163
$\beta_{\mathrm{Pz700{\sim}DIFFICULTY}}$							
Atomoxetine	-1.30	0.29	-1.76	0.30	F	5.65	.024*
Placebo	-1.27	0.22	-1.91	0.27	T	0.08	.775
					F x T	0.24	.630
$\beta_{\mathrm{Pz700{\sim}NOISE}}$							
Atomoxetine	0.15	0.24	0.43	0.18	F	9.85	.004**
Placebo	-0.28	0.19	0.84	0.22	T	0.01	.941
					F x T	3.73	.064

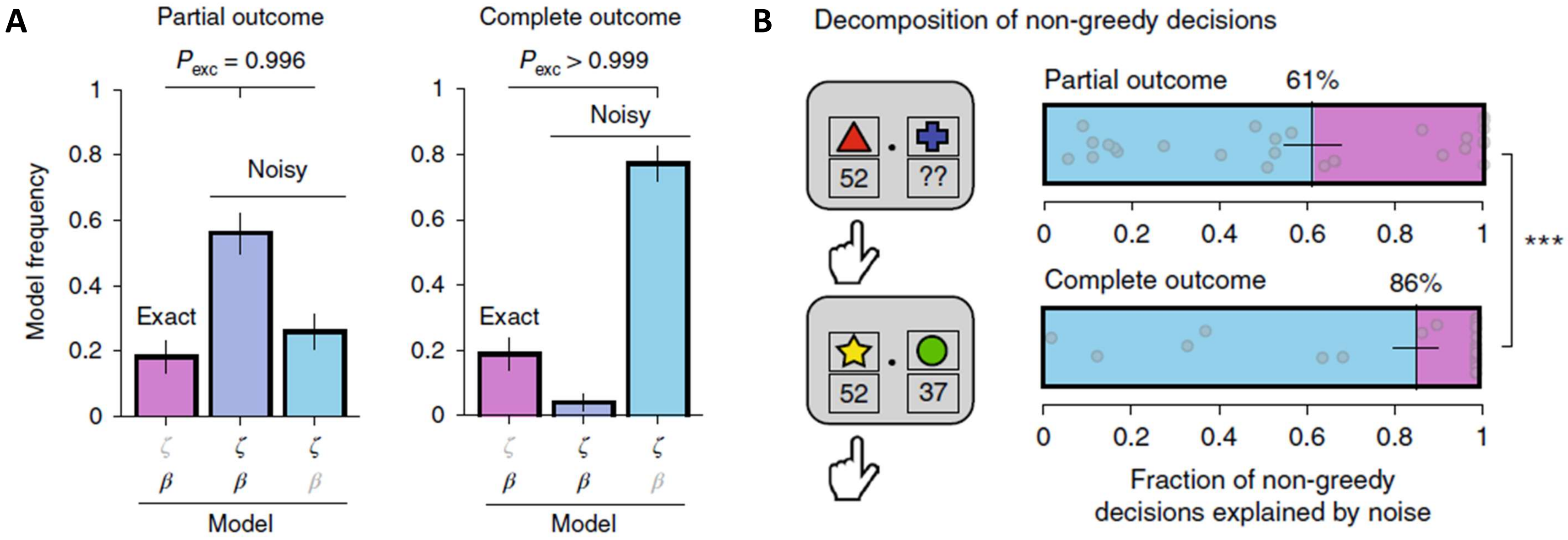
Note. In a first step, we examined for each participant (*N* = 30) the impact of the RL model variables (RPE, difficulty, noise) on the single-trial parietal activity within the time window of valence P700 with separate regression analyses for the four conditions (feedback: partial vs. complete; treatment: atomoxetine vs. placebo). In a second step, we ran a mixed-effects ANOVA to examine the effect of feedback and treatment condition on the standardised regression coefficients (assuming equal variances across conditions, however, non-normality of the distributions). The Greenhouse-Geisser (1959) correction did not yield different *p* values. RPE = reward prediction error; ANOVA = analysis of variance; Partial = partial feedback condition; Complete = complete feedback condition; F = feedback; T = treatment. For illustration of the findings, see Figure 8.

*p < .05 **p < .01.

Figures

Figure 1

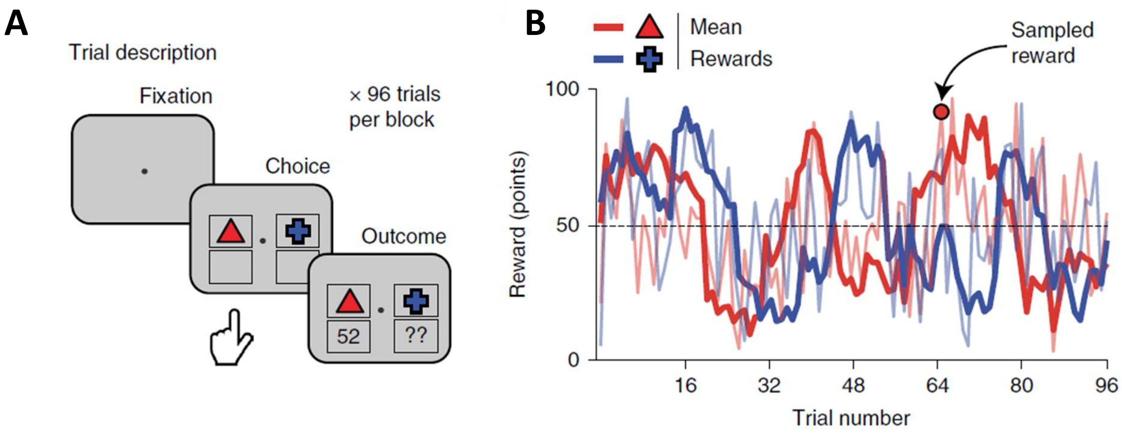
Original Findings of Findling et al.'s (2019) Study of the Model Selection Procedure and the Decomposition of Non-Greedy Decisions



Note. Panel A: Results of the Bayesian model selection (BMS) procedure (pooled over two experiments, $N = 59$) by Findling et al. (2019), showing the best model fit for the noisy RL softmax model (entailing the learning imprecision ζ and exploration parameter - inverse temperature - $\beta = 1/\tau$) for the partial feedback condition (left), and the best fit for the noisy RL argmax model (only entailing the learning imprecision parameter ζ) for the complete feedback condition (right). P_{exc} = exceedance probability. Panel B: Results of the Decomposition of non-greedy decisions by Findling et al. (2019), showing the fraction of non-greedy decisions explained by learning imprecision (termed as “noise”; blue area) is lower in the partial (top; 61%) than in the complete feedback condition (bottom: 86%). The purple area reflects the fraction of non-greedy decisions explained by choice stochasticity. From “Computational Noise in Reward-Guided Learning Drives Behavioral Variability in Volatile Environments,” by C. Findling, V. Skvortsova, R. Dromnelle, S. Palminteri, and V. Wyart, 2019, *Nature Neuroscience*, *22*(12), p. 2068 (https://doi.org/10.1038/s41593-019-0518-9). Copyright 2019 by Springer Nature America, Inc.

Figure 2

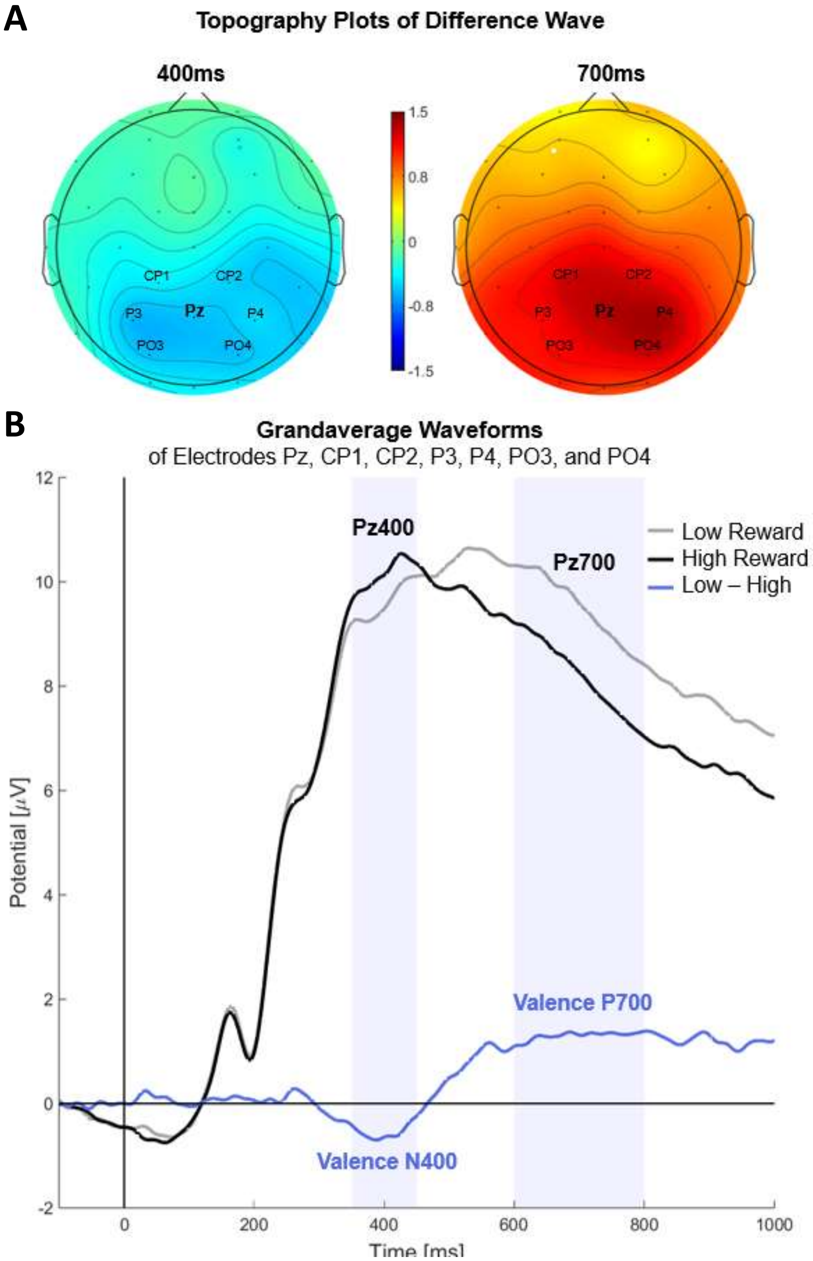
Trial Description and Visualisation of the Mean and Sampled Rewards Across Trials



Note. The number of trials differed in our experiment: 48 trials in the two training blocks, and 56 trials in the twelve experimental blocks. Panel A: In every trial, participants chose between two differently coloured shapes and observed the associated reward (1-99 points), which was later converted into real financial incentives. Panel B: Rewards were sampled (bright lines) from distributions with means drifting independently across trials (thick lines). From "Computational Noise in Reward-Guided Learning Drives Behavioral Variability in Volatile Environments," by C. Findling, V. Skvortsova, R. Dromnelle, S. Palminteri, and V. Wyart, 2019, *Nature Neuroscience*, *22*(12), p. 2067 (https://doi.org/10.1038/s41593-019-0518-9). Copyright 2019 by Springer Nature America, Inc.

Figure 3

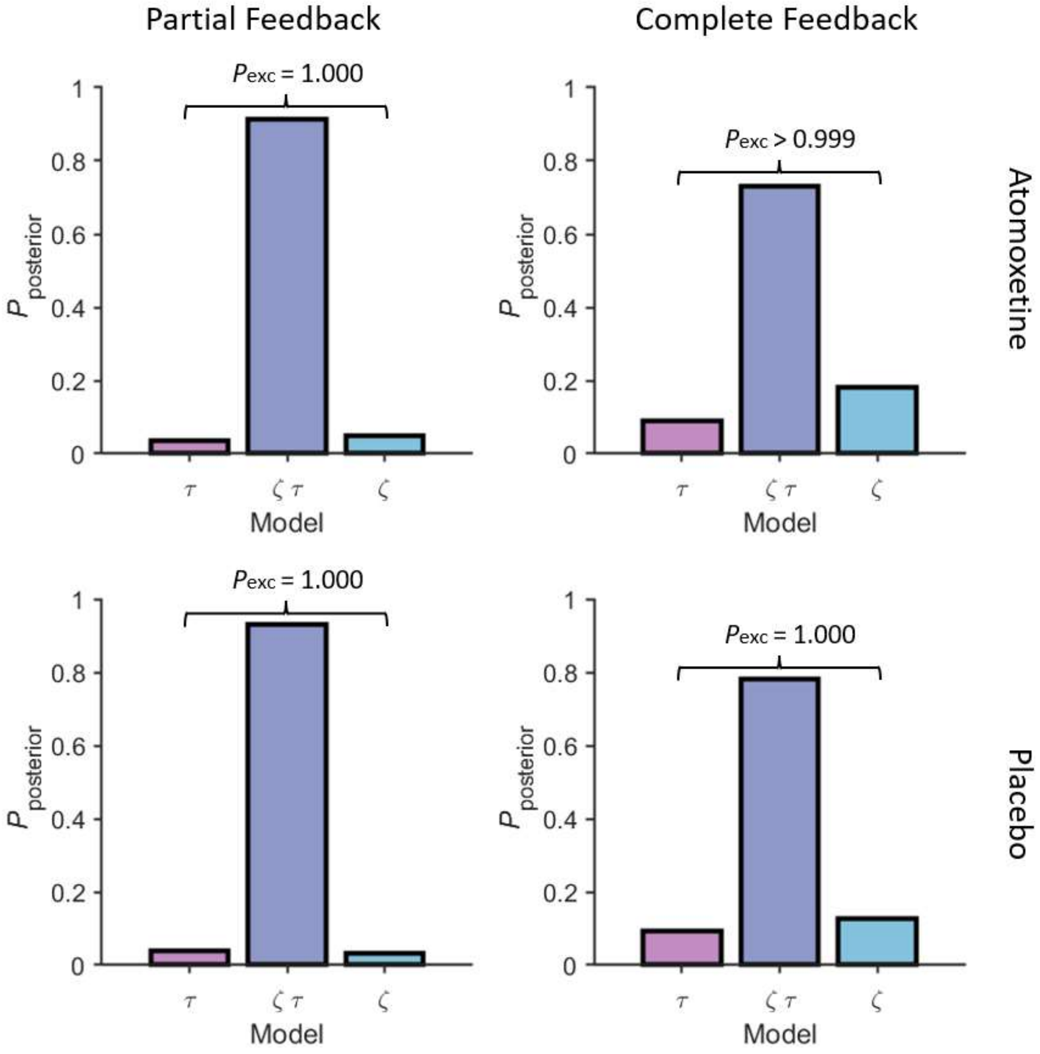
Spatial and Temporal Occurrence of the Valence-Sensitive ERP Components



Note. The topography plots of the reward valence difference wave (low reward indicating negative valence minus high reward indicating positive valence) plotted for each 50 ms interval (from feedback onset to 1000 ms post-feedback) revealed a maximum negative and positive deflection at parietal site (electrode Pz and the neighbouring electrodes CP1, CP2, P3, P4, PO3, and PO4) at around 400 and 700 ms after feedback onset. Panel A: Topography plot of the grandaverage difference wave of reward valence averaged across the parietal cluster only, showing a pronounced negativity at 400 ms (left) and a pronounced positivity at 700 ms (right). Panel B: Grandaverage waveforms of low reward trials (≤ 50 points; grey line) and high reward trials (> 50 points; black line), and the difference wave thereof (low minus high reward trials; blue line). The blue area reflects the time-windows we identified to enclose peak activity of the components (valence N400: 350 to 450 ms; valence P700: 600 to 800 ms).

Figure 4

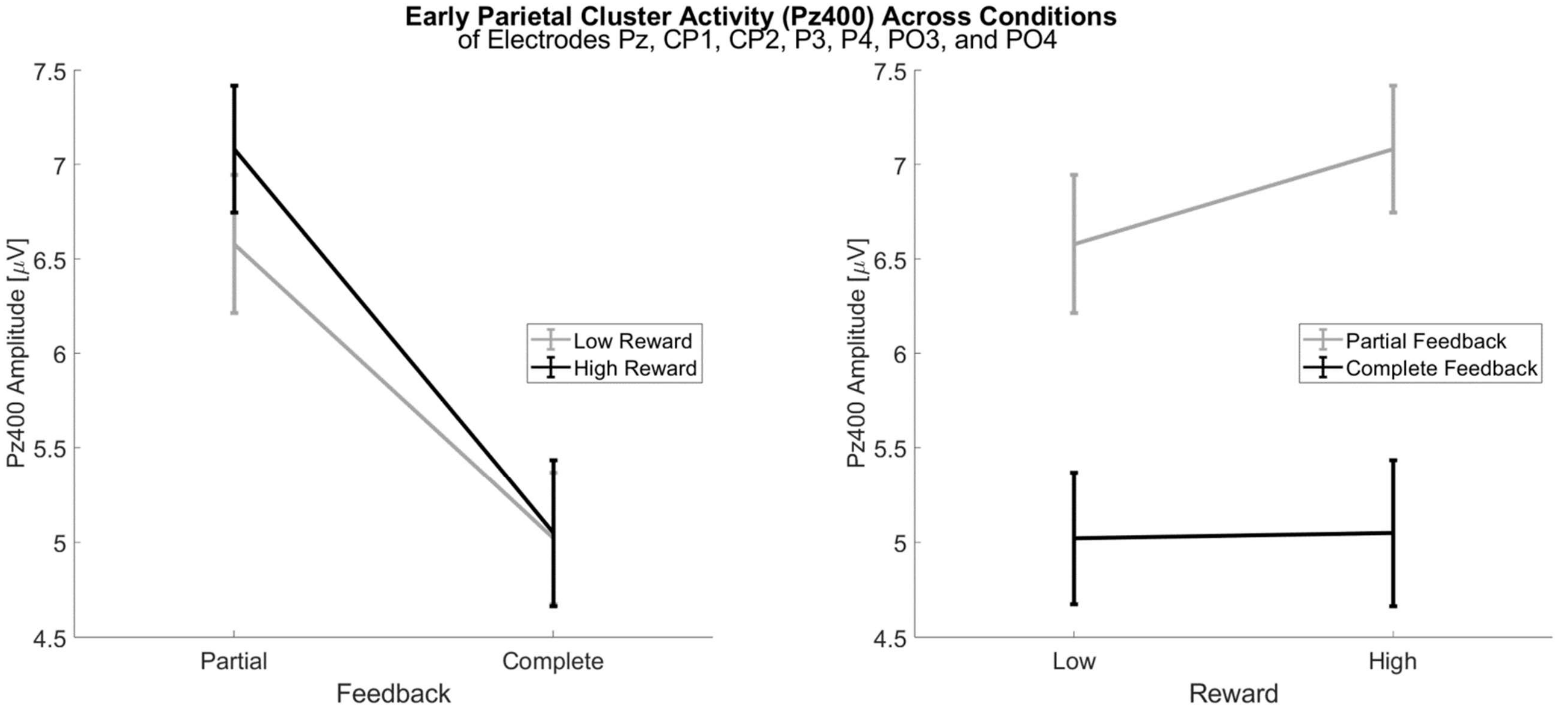
Model Comparison of the RL Models With and Without Exploration and Learning Imprecision Parameters



Note. Results of the Bayesian model selection (BMS) procedure ($N = 30$), showing the best model fit for the noisy RL softmax model (entailing the learning imprecision ζ and exploration parameter τ) for the partial feedback (left), the complete feedback (right), as well as for both treatment conditions (top: atomoxetine; bottom: placebo). The posterior probability and the exceedance probability are model comparison metrics, indicating how likely the model is given the data (posterior probability), respectively how more likely the model is than any of the other models given the data (exceedance probability). P_{exc} = exceedance probability; $P_{posterior}$ = posterior probability.

Figure 5

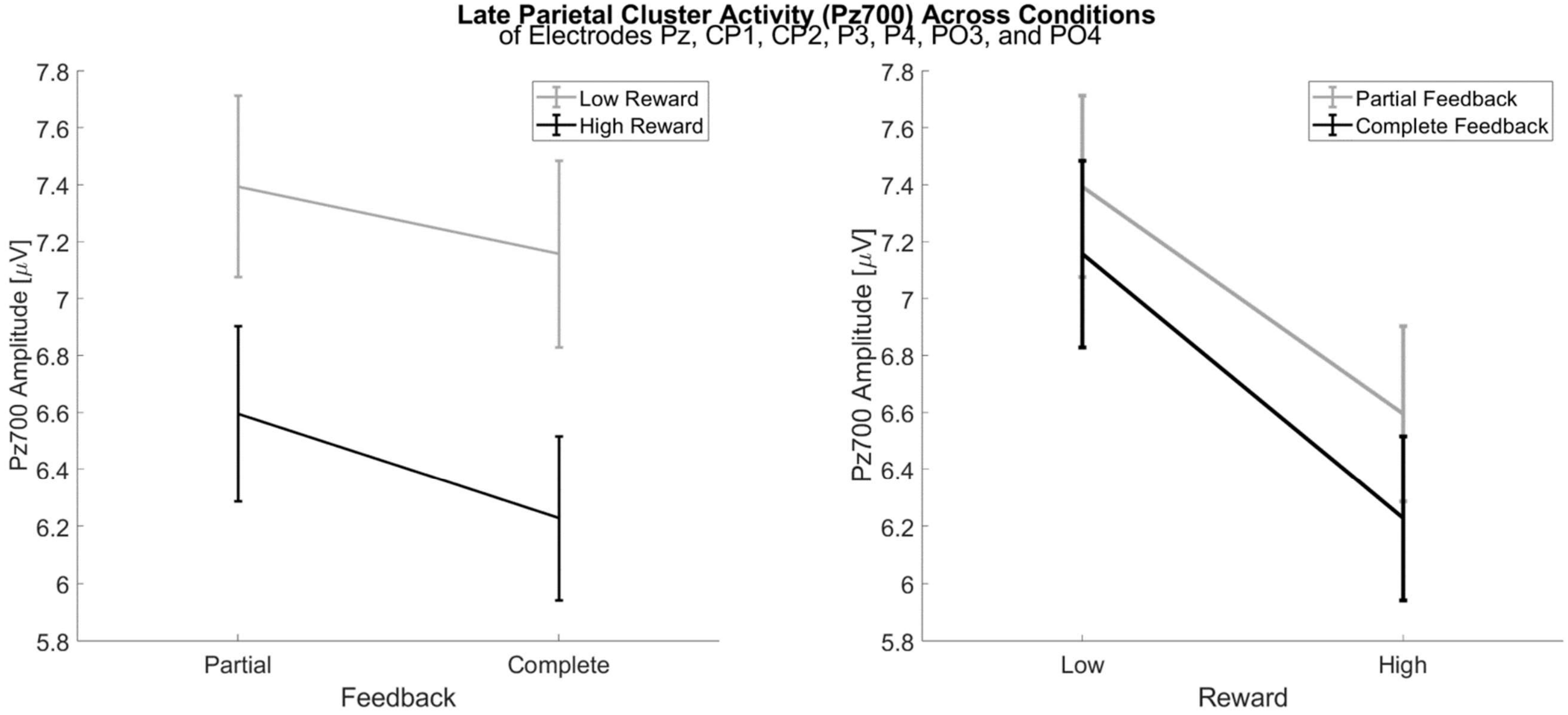
Mean Pz400 Amplitude Across the Model-Free Conditions Feedback Type and Reward Magnitude



Note. Average amplitude of the early parietal cluster between 350 and 450 ms after feedback onset of electrodes Pz, CP1, CP2, P3, P4, PO3, and PO4 across the feedback (partial vs. complete) and reward (low: ≤ 50 points vs. high: > 50 points) conditions. Error bars represent standard errors. The Pz400 amplitude served as dependent variable of the repeated measures ANOVA with feedback type and reward magnitude as independent variables, resulting in a significant main effect of feedback type (left), but not reward magnitude (right).

Figure 6

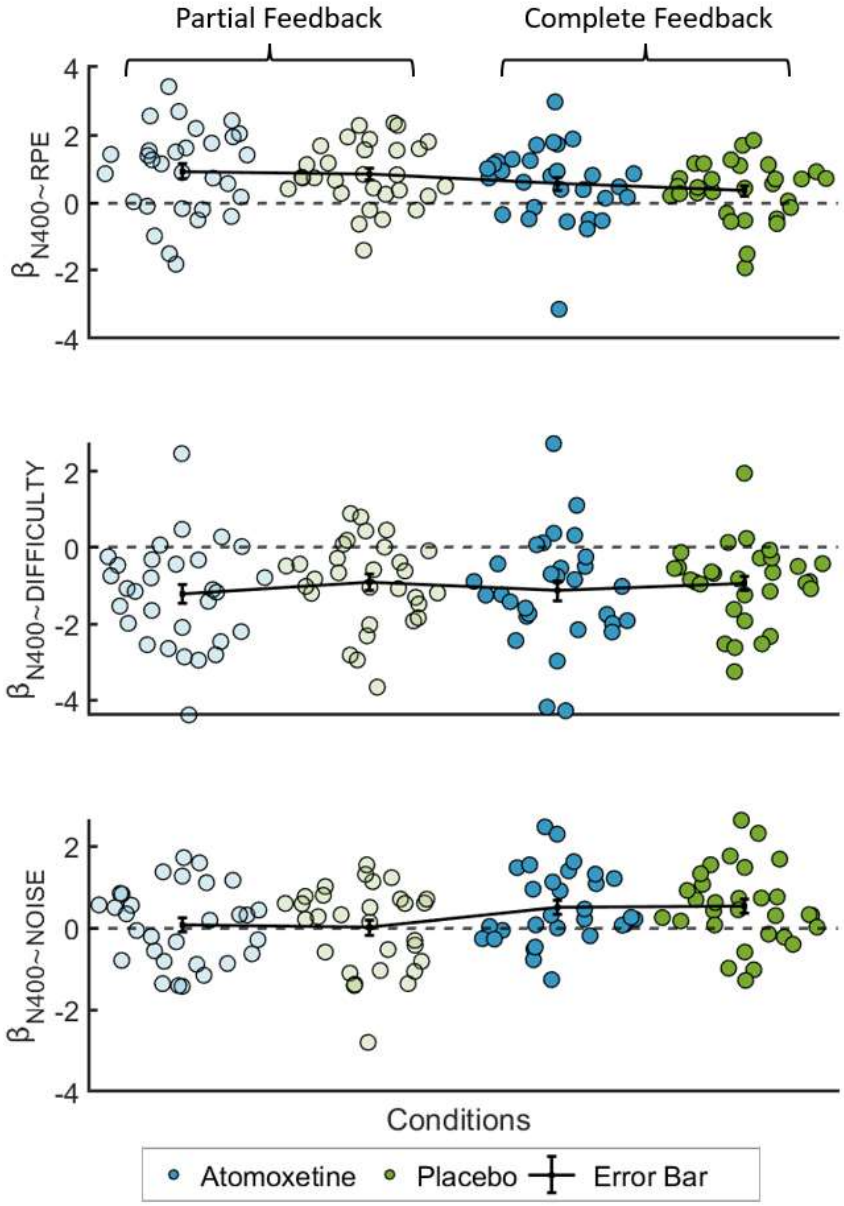
Mean Pz700 Amplitude Across the Model-Free Conditions Feedback Type and Reward Magnitude



Note. Average amplitude of the late parietal cluster between 600 and 800 ms after feedback onset of electrodes Pz, CP1, CP2, P3, P4, PO3, and PO4 across the feedback (partial vs. complete) and reward (low: ≤ 50 points vs. high: > 50 points) conditions. Error bars represent standard errors. The Pz700 amplitude served as dependent variable of the repeated measures ANOVA with feedback type and reward magnitude as independent variables, resulting in a significant main effect of reward magnitude (right), but not feedback type (left).

Figure 7

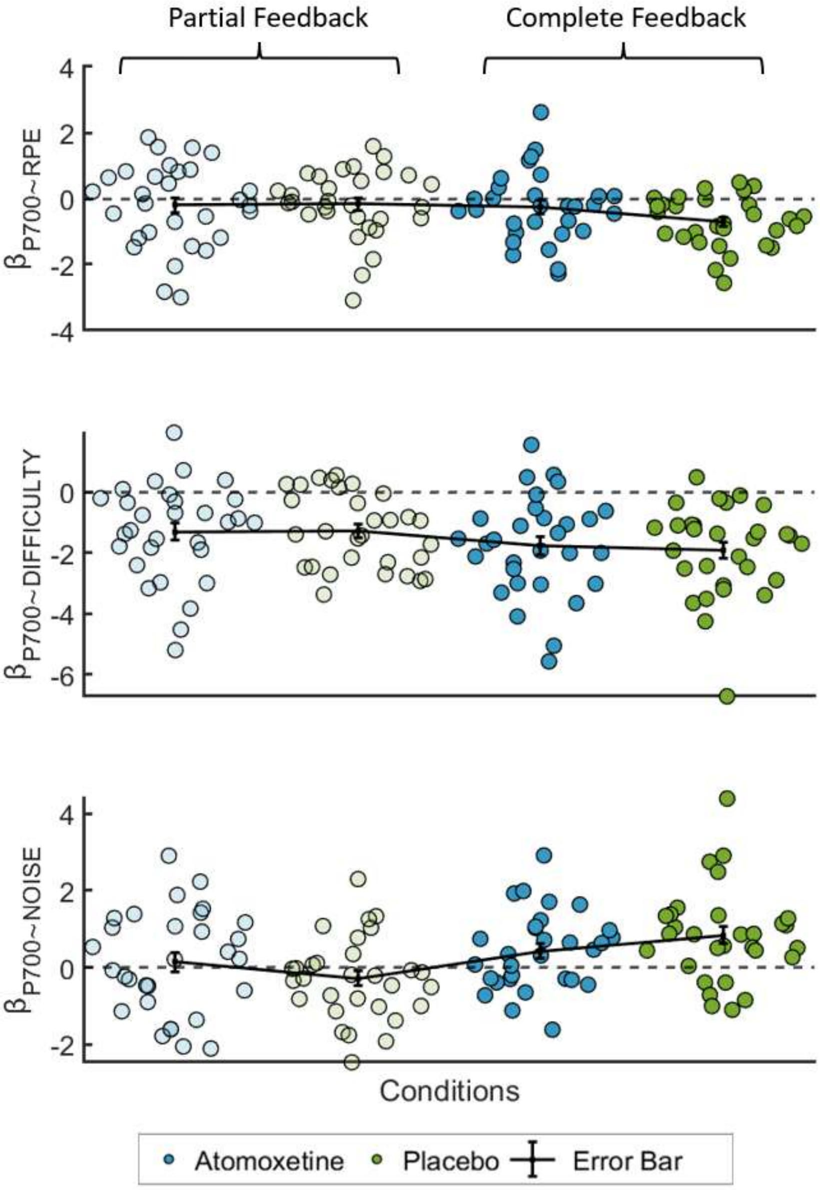
Standardised Regression Coefficients for Parietal EEG Activity Within the Time Window of Valence N400 Across Conditions (Treatment x Feedback)



Note. Standardised regression coefficients of reward prediction error (RPE; top), choice difficulty (middle), and learning noise (bottom) for the averaged single-trial EEG activity of the parietal clusters (Pz, CP1, CP2, P3, P4, PO3, and PO4) within the time window of valence N400 (350 to 450 ms after feedback) for each participant plotted across conditions. The light marker colours refer to the partial feedback condition (left), while the dark colours refer to the complete feedback condition (right). Atomoxetine = blue colour; Placebo = green colour. Error bars represent the mean and standard errors.

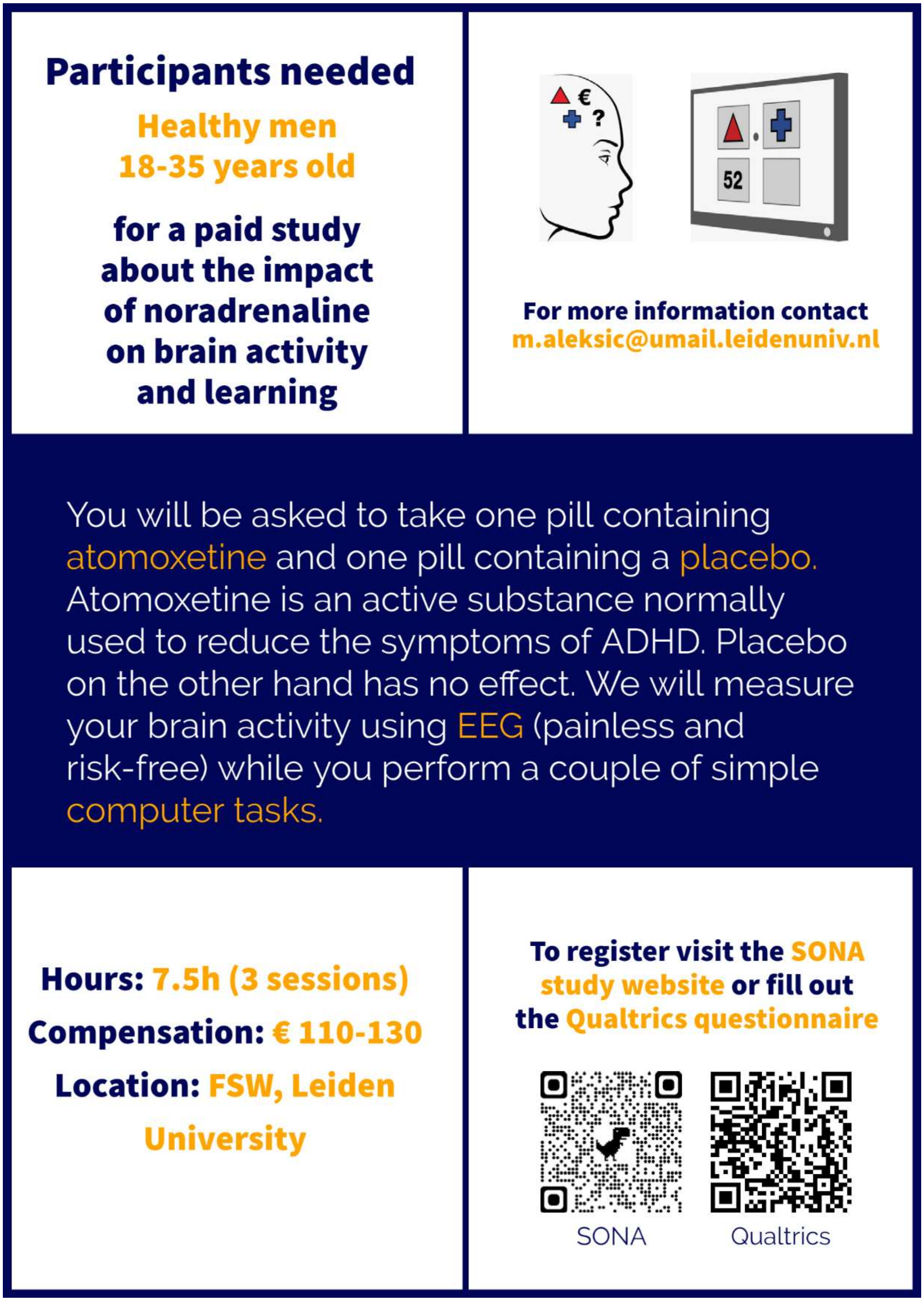
Figure 8

Standardised Regression Coefficients for Parietal EEG Activity Within the Time Window of Valence P700 Across Conditions (Treatment x Feedback)



Note. Standardised regression coefficients of reward prediction error (RPE; top), choice difficulty (middle), and learning noise (bottom) for the averaged single-trial EEG activity of the parietal clusters (Pz, CP1, CP2, P3, P4, PO3, and PO4) within the time window of valence P700 (600 to 800 ms after feedback) for each participant plotted across conditions. The light marker colours refer to the partial feedback condition (left), while the dark colours refer to the complete feedback condition (right). Atomoxetine = blue colour; Placebo = green colour. Error bars represent the mean and standard errors.

Appendix A



Participants needed

Healthy men
18-35 years old

for a paid study
about the impact
of noradrenaline
on brain activity
and learning

For more information contact
m.aleksic@umail.leidenuniv.nl

You will be asked to take one pill containing atomoxetine and one pill containing a placebo. Atomoxetine is an active substance normally used to reduce the symptoms of ADHD. Placebo on the other hand has no effect. We will measure your brain activity using EEG (painless and risk-free) while you perform a couple of simple computer tasks.

Hours: 7.5h (3 sessions)

Compensation: € 110-130

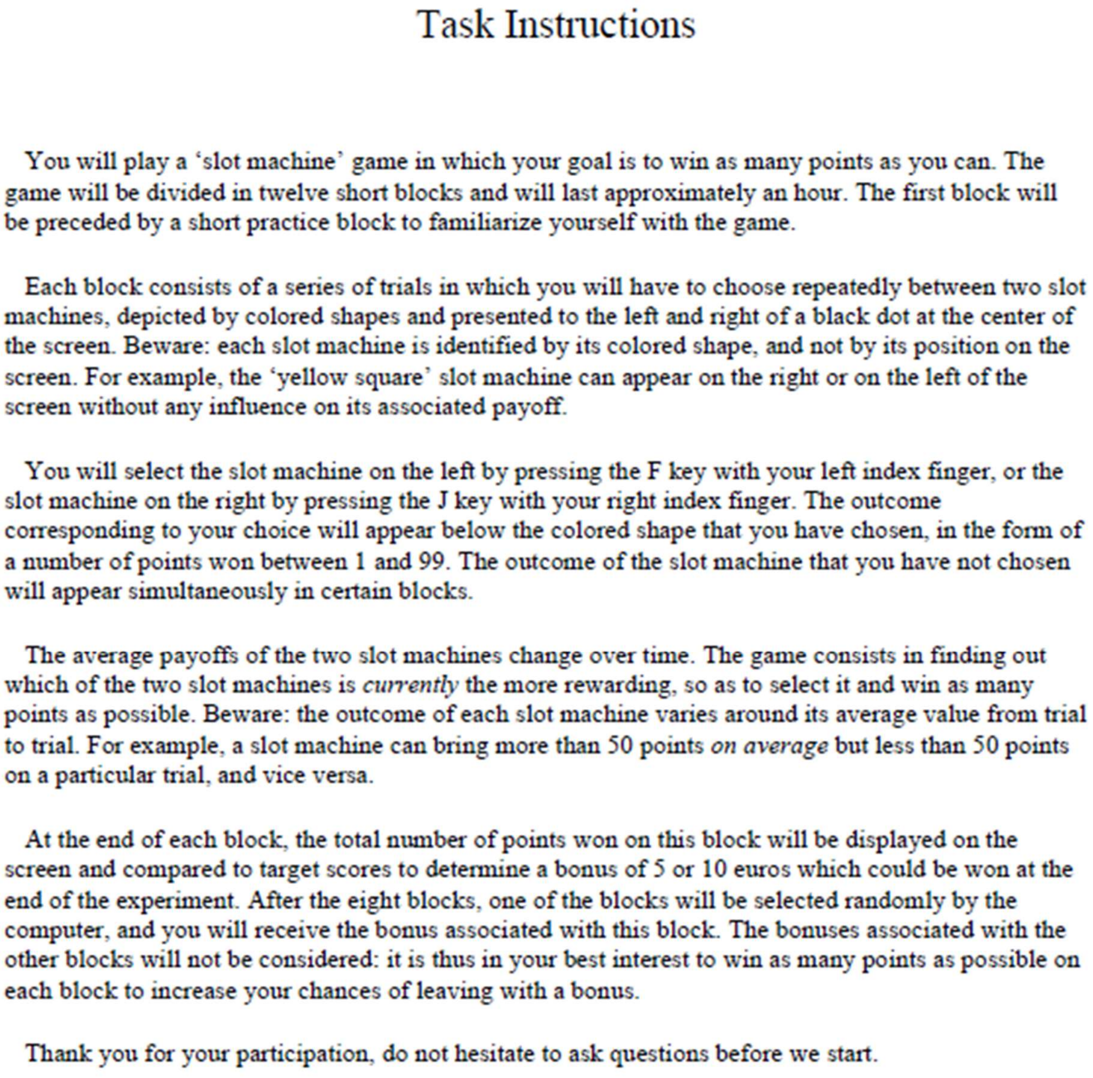
Location: FSW, Leiden University

To register visit the SONA study website or fill out the Qualtrics questionnaire

SONA

Qualtrics

Appendix B



Task Instructions

You will play a 'slot machine' game in which your goal is to win as many points as you can. The game will be divided in twelve short blocks and will last approximately an hour. The first block will be preceded by a short practice block to familiarize yourself with the game.

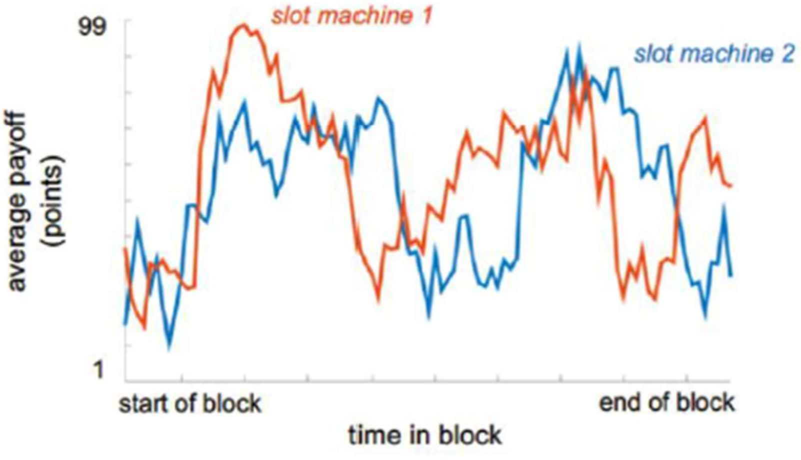
Each block consists of a series of trials in which you will have to choose repeatedly between two slot machines, depicted by colored shapes and presented to the left and right of a black dot at the center of the screen. Beware: each slot machine is identified by its colored shape, and not by its position on the screen. For example, the 'yellow square' slot machine can appear on the right or on the left of the screen without any influence on its associated payoff.

You will select the slot machine on the left by pressing the F key with your left index finger, or the slot machine on the right by pressing the J key with your right index finger. The outcome corresponding to your choice will appear below the colored shape that you have chosen, in the form of a number of points won between 1 and 99. The outcome of the slot machine that you have not chosen will appear simultaneously in certain blocks.

The average payoffs of the two slot machines change over time. The game consists in finding out which of the two slot machines is *currently* the more rewarding, so as to select it and win as many points as possible. Beware: the outcome of each slot machine varies around its average value from trial to trial. For example, a slot machine can bring more than 50 points *on average* but less than 50 points on a particular trial, and vice versa.

At the end of each block, the total number of points won on this block will be displayed on the screen and compared to target scores to determine a bonus of 5 or 10 euros which could be won at the end of the experiment. After the eight blocks, one of the blocks will be selected randomly by the computer, and you will receive the bonus associated with this block. The bonuses associated with the other blocks will not be considered: it is thus in your best interest to win as many points as possible on each block to increase your chances of leaving with a bonus.

Thank you for your participation, do not hesitate to ask questions before we start.


slot machine 1
slot machine 2
average payoff (points)
start of block
end of block
time in block