Cognitive, Affective, & Behavioral Neuroscience (2022) 22:1108–1129
https://doi.org/10.3758/s13415-022-00996-z

RESEARCH ARTICLE

Pupil dilation and response slowing distinguish deliberate explorative choices in the probabilistic learning task

Galina L. Kozunova[1] · Ksenia E. Sayfulina[1] · Andrey O. Prokofyev[1] · Vladimir A. Medvedev[1] · Anna M. Rytikova[1] · Tatiana A. Stroganova[1] · Boris V. Chernyshev[1]

Accepted: 7 March 2022 / Published online: 1 April 2022

Abstract

This study examined whether pupil size and response time would distinguish directed exploration from random exploration and exploitation. Eighty-nine participants performed the two-choice probabilistic learning task while their pupil size and response time were continuously recorded. Using LMM analysis, we estimated differences in the pupil size and response time between the advantageous and disadvantageous choices as a function of learning success, i.e., whether or not a participant has learned the probabilistic contingency between choices and their outcomes. We proposed that before a true value of each choice became known to a decision-maker, both advantageous and disadvantageous choices represented a random exploration of the two options with an equally uncertain outcome, whereas the same choices after learning manifested exploitation and direct exploration strategies, respectively. We found that disadvantageous choices were associated with increases both in response time and pupil size, but only after the participants had learned the choice-reward contingencies. For the pupil size, this effect was strongly amplified for those disadvantageous choices that immediately followed gains as compared to losses in the preceding choice. Pupil size modulations were evident during the behavioral choice rather than during the pretrial baseline. These findings suggest that occasional disadvantageous choices, which violate the acquired internal utility model, represent directed exploration. This exploratory strategy shifts choice priorities in favor of information seeking and its autonomic and behavioral concomitants are mainly driven by the conflict between the behavioral plan of the intended exploratory choice and its strong alternative, which has already proven to be more rewarding.

Keywords Pupil size · Response time · Direct and random exploration · Probability learning · Conflict

"A bird in the hand is worth two in the bush" – will you follow this common wisdom, or will you ever abandon something good you already have and venture into the unknown in the vague hope of a bigger win? In a probabilistic environment, people usually tend to imagine hidden regularities in the outcomes of their actions, even when no such regularities actually exist (Ellerby & Tunney, 2017; Unturbe & Corominas, 2007). Attempting to test their surmises and catch a lucky break, people explore apparently disadvantageous options instead of just sticking to familiar profitable ones (Shanks et al., 2002). In fact, by doing so in probabilistic experimental tasks involving truly random and mutually independent choice outcomes, they actually fail to maximize their profits (Guttel & Harel, 2005; Unturbe & Corominas, 2007; Vulkan, 2000).

In contrast to typical artificial experimental conditions, outcomes of one's actions in real life may be nonrandom and interdependent, e.g., the outcome of the next trial may be a consequence of the outcome of the previous trial. In this respect, exploring uncertain options instead of continuously exploiting a more rewarding alternative might bring new information about possible rewards, and thus might increase payoffs in the long run (Cogliati Dezza et al., 2017; Sayfulina et al., 2020). From this perspective, the occasional switches from the prepotent value-driven advantageous response tendency to a disadvantageous choice in the probabilistic task may be considered as directed exploration—exploratory behavior that occurs when our desire for information overrides our need for reward. This crucial distinction between two qualitatively different types of

✉ Boris V. Chernyshev
b_chernysh@mail.ru

[1] Center for Neurocognitive Research (MEG-Center), Moscow State University of Psychology and Education, 29 Sretenka str, Moscow 127051, Russia

exploration—directed and random exploration—has recently gained growing attention in the literature (Payzan-LeNestour & Bossaerts, 2012; Schulz & Gershman, 2019; Schwartenbeck et al., 2019; Wilson et al., 2014; Zajkowski et al., 2017). Directed exploration is intentional and specifically related to information-seeking targeted at the most uncertain option. On the contrary, random exploration is essentially a noisy response-generation process, leading to choices made by chance. Importantly, such type of exploration may be observed during early stages of reinforcement learning (Averbeck, 2015; Cogliati Dezza et al., 2017; Schulz & Gershman, 2019; Sutton & Barto, 1999), when all options are uncertain for the subject, or in conditions when the value of the most valuable option has decreased (Schwartenbeck et al., 2019)—thus also creating a need for learning the new rule of choice-reward contingency.

Random and directed forms of exploration are supposed to differ in underlying brain mechanisms (Warren et al., 2017; Zajkowski et al., 2017). Yet, the distinction between the two forms of exploration has not been accounted for in many previous physiological studies of exploration–exploitation dilemma.

For more than a decade, the adaptive gain theory (Aston-Jones & Cohen, 2005) has been an important theoretical background that guided many neurocognitive studies of the balance between exploitation and exploration. According to this theory, the locus coeruleus-noradrenergic (LC-NA) neuromodulatory system plays an essential role in regulating the balance between the two strategies (Aston-Jones & Cohen, 2005; Usher et al., 1999). Specifically, this theory proposes that the tonic mode of the LC-NA neuromodulatory system promotes disengagement from the task and processing of task-irrelevant stimuli, thus creating a proper control state for exploring the new options. The increased neuromodulatory activity occurring over rather long timescale is accompanied by tonic pupil dilation, which is measured during pre-trial baseline, and it is believed to be a reliable proxy of LC-NA activation in the brain (Joshi & Gold, 2020). The tonic pupil dilation has been widely used to investigate the physiological basis of exploratory behavior (Gilzenrat et al., 2010; Jepma & Nieuwenhuis, 2011; Jepma et al., 2010).

In line with the concept of task disengagement central to the adaptive gain theory, many previous pupil studies of exploration–exploitation trade-off used behavioral tasks that explicitly forced participants to commit exploratory choices: typically that was achieved by decreasing the reward associated with the preferred choice (Daw et al., 2006; Jepma & Nieuwenhuis, 2011; Payzan-LeNestour & Bossaerts, 2012). Such exploration related to task disengagement has important hallmarks of random exploration (Wilson et al., 2021), e.g., as operationalized in Schwartenbeck et al. (2019). Yet little is known about the neurocognitive mechanisms associated with self-generated directed exploration targeted at information seeking (Zajkowski et al., 2017). In sharp contrast with random exploration, decision making during directed exploration focuses on active processing of choice-related information, thus emphasizing the deliberative process that requires the subject's attention. Pupil size, as a component of phasic arousal, changes rapidly in response to cognitive operations underlying such attentional decision making (Poe et al., 2020). Specifically, recent research has shown that neural encoding of uncertainty or conflict can trigger fast task-evoked pupil dilation within the same trial in probabilistic reinforcement learning tasks (Van Slooten et al., 2018), as well as in a target discrimination task (Gilzenrat et al., 2010). This suggests that the decision to deliberately choose an option with a more uncertain outcome is associated with phasic rather than tonic pupil-related arousal.

In the current study, we investigated the observable physiological (pupil size) and behavioral (response time) concomitants of self-generated directed exploratory choices initiated by a participant on his or her own, i.e., in the absence of any external triggers and within a uniform series of trials under unchanging reward probabilities. As far as we know, the pupil-related arousal during directed exploration has never been addressed in the literature related to the exploration–exploitation dilemma.

We used a two-alternative repetitive choice task: with one option bringing more gains than losses, and the other one bringing more losses than gains. Participants were learning the task rules by trial and error. We primarily aimed to explore the distinction between the low-payoff (LP) and high-payoff (HP) choices, which were committed after a participant had learnt the choice-reward contingency. At this stage, participants had already acquired statistics of reward values associated with choice option, i.e., they possessed an internal utility model that guided them to prefer the HP option to the LP one. We reasoned that when choosing the option yielding a lesser (mathematical) expectation of the reward, people were engaged in an active effort to gather the information they are interested in (directed exploration), and they deliberately violated the acquired utility model.

To check this assumption, we contrasted LP choices made after successful learning to LP choices committed during the experimental blocks, in which participants failed to acquire a preference for the advantageous option, and thus did not possess any internal utility model relevant to the task rules.

Our basic predictions were related to the conflict/uncertainty that specifically characterizes the suboptimal LP choices violating the internal utility model. Indeed, there is a substantial evidence that to explore and gather information on a less certain, risky, and potentially nonrewarding option, participants have to inhibit the tendency to choose a highly rewarded safe alternative (Cogliati Dezza et al., 2017; Daw et al., 2006). Assuming that the drive to commit the exploratory choice of a disadvantageous low-payoff (LP) option can be induced internally as a matter of actively probing the environment, we can predict that the decision to commit a directed exploration would be accompanied by a conflict between seeking information concerning the uncertain options and seeking greater immediate profit associated with the other options as predicted by the internal utility model of the task.

On the basis of our hypothesis and previous pupil and response time (RT) studies of decision making (Cavanagh et al., 2014; Egner, 2007; Gilzenrat et al., 2010; Hershman & Henik, 2019; Laeng et al., 2011; Lin et al., 2018; Satterthwaite et al., 2007; Urai et al., 2017; Van Slooten et al., 2018), we anticipated that conflict/uncertainty pertaining to self-generated exploratory choices would lead to greater phasic pupil-related arousal and RT slowing during the LP compared with “safe” HP choices, but only when the decision-maker has learnt the choice-reward contingencies. We also expected that greater pupil dilation and decision costs would differentiate the LP choices made in the context of directed exploration from “random” LP choices committed under “no learning” condition. Since tonic LC-NA activation measured via pretrial pupil dilation has been implicated in exploratory behaviour (Gilzenrat et al., 2010; Jepma & Nieuwenhuis, 2011), we also investigated whether similar effects exist during self-generated directed exploratory behavior.

Methods

Participants

Ninety-four volunteers recruited from the community participated in the experiment (46 men and 48 women), aged 25.9 ± 5.6 years ($M \pm SD$). All participants reported no neurological disorders or severe visual impairments; visual acuity was within ± 2.5 diopters, at least for a better-seeing eye. During the experiment, the participants did not use any vision correction devices (such as glasses or contact lenses).

The study was conducted following the ethical principles regarding human experimentation (Helsinki Declaration) and approved by the Ethics Committee of the Moscow State University of Psychology and Education. All participants signed the informed consent before the experiment.

Procedure

During the experiment, participants were comfortably seated in an armchair and placed their heads on the chin rest to minimize involuntary head movements. We used the modified probabilistic learning task (Frank et al., 2004; Kozunova et al., 2018), which was rendered as a computer game. On each trial, participants had to choose between two stimuli presented on the screen simultaneously. One stimulus was assigned as the advantageous (choosing this stimulus led to gains on 70% of trials and to losses on 30% of trials) and the other one as the disadvantageous (leading to gains on 30% of trials and to losses on 70% of trials). Probabilities were kept constant and did not change during the experiment. Losses and gains were assigned by a computer in a quasi-random order. Before the experiment, the participants were informed that the two stimuli were not equal in terms of the number of gains they could bring; yet no further information was revealed to the participants. Thus, the instruction could not prompt any specific choice strategies, and the participants had to learn from their own experience in trial-and-error fashion.

Each pair of stimuli comprised two images of the same Hiragana hieroglyph rotated at two different angles and rendered in white-on-black background (Fig. 1). The size of the stimuli was 1.54 × 1.44°, which well fits into the fovea area. The stimuli were equalized in size, brightness, perceptual complexity, and spatial position. The two stimuli were located symmetrically on the left and on the right of the screen at 1.5° to the left and to the right from the screen center. The location of stimuli was alternated pseudorandomly from trial to trial with equal probability.

Before the start of the experiment, participants completed a quick test for visual discrimination and recognition of figures similar to those used during the experiment. All participants passed the test well.

On each trial, before the stimulus onset, a white fixation cross on black background was presented for 150 ms (Fig. 1). Stimuli remained on the screen until a button was pressed by the participant; the instruction given to the participants did not require keeping gaze fixation during stimulus presentation, and the participants were allowed to freely view the stimuli. Aiming to avoid time pressure on participants and thus minimize the number of impulsive decisions, we did not impose any time limit on the response time.

During the experiments, participants continuously kept their index and middle fingers of the right hand on two buttons of the response pad (Current Designs, Philadelphia, Inc., PA). To choose one of the two stimuli, participants pressed one of the buttons according to the location of the chosen stimulus on the screen (i.e., they pressed the left button to choose the left stimulus and the right button to choose the right one).

Immediately after the button press, the screen was cleared and remained empty (black). After a delay of 1 s following the button press, the visual feedback was presented for 500 ms. The feedback informed the participants about the number of points they received or lost on the current trial (Fig. 1). These points were accumulated throughout the experiment; at the end of each block, participants were

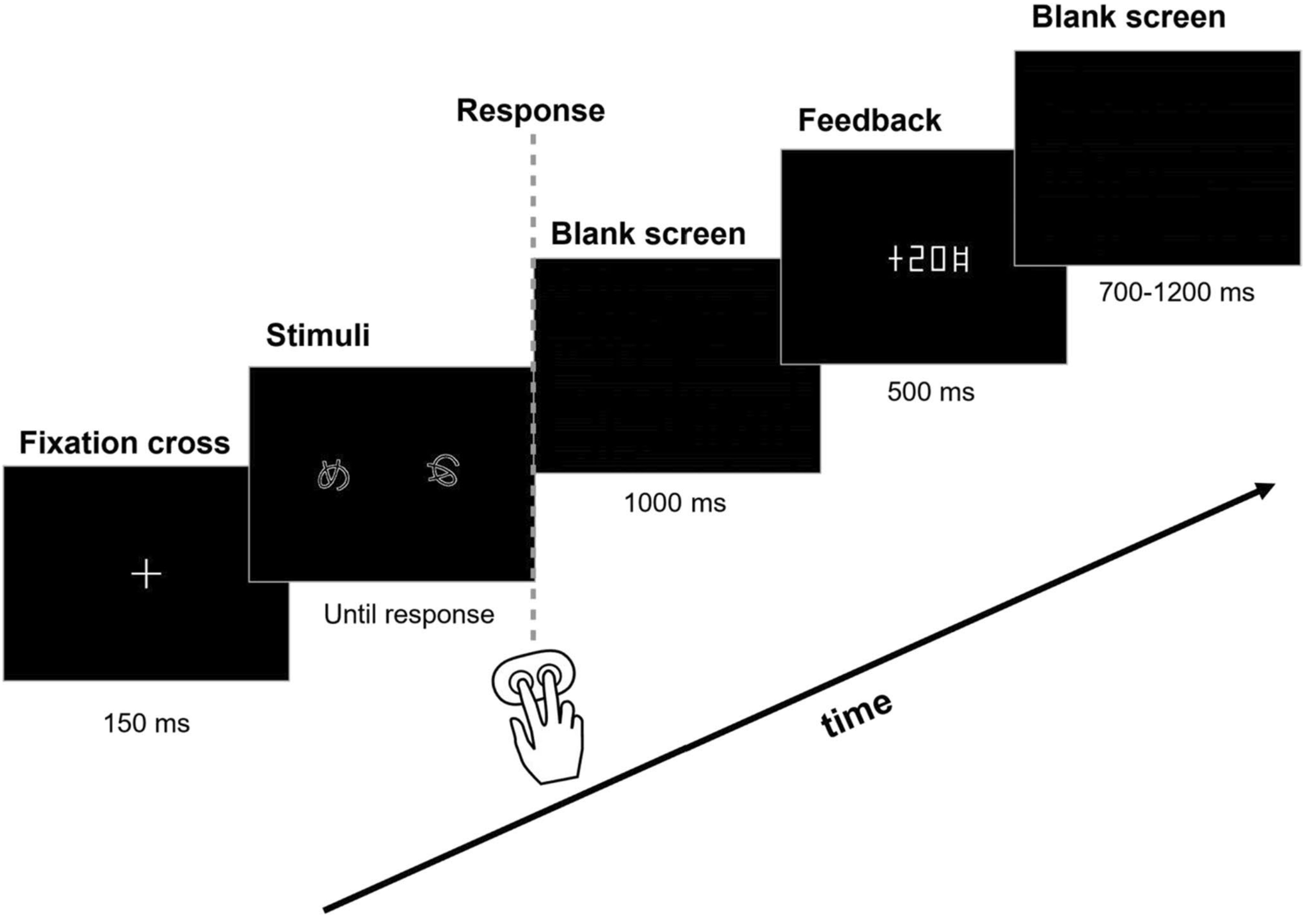

Fig. 1 Probabilistic value-based decision-making task. Each stimulus pair (a Hiragana hieroglyph rotated at two different angles) was presented repeatedly, and the left–right position of the stimulus on the screen was counterbalanced throughout an experimental block. Stimulus pairs and the expected value of each choice in a pair varied between the five blocks. Participants learned to select the more profitable of the two options (the advantageous one) solely from probabilistic feedback. After each trial, the number of points earned was displayed after a 1,000-ms delay after a participant made a choice of one of the options by pressing the left or right button (for the left and right choices respectively). See text for details

shown their total accumulated score. The screen was black during the intertrial interval.

We used an intertrial interval ranging from 700 to 1,400 ms, which varied in a quasi-random order (flat distribution). We used a rather short intertrial interval, thus keeping the duration of the whole experiment to a minimum. This was done with the aim of preventing fatigue and boredom in participants and maintaining their interest and motivation. This intertrial interval was shorter than the duration of the pupillometric effects; we addressed this issue at the stage of preprocessing the pupillometric data (see below). The intertrial interval was varied in order to prevent rhythmical responding that could lead to impulsive or perseverative responses.

The experiment involved five similar blocks. Each block included 40 trials and lasted approximately 5 min. A short rest lasting approximately 1 min was introduced between blocks. The total duration of the experiment was approximately 30 min. A new stimulus pair was used in each block. Although the probabilities of gains and losses associated with the advantageous and disadvantageous stimuli were kept constant throughout the whole experiment, blocks differed in the exact number of points assigned as gains and losses. We used five reinforcement schemes with the following magnitudes of gains and losses, respectively: (I) + 20 & 0, (II) 0 & − 20, (III) + 50 & + 20, (IV) − 20 & − 50, (V) + 20 & − 20. Changing reinforcement scale between blocks was needed to make the blocks appear less similar and thus to make the participants learn during each block anew.

The order of reinforcement schemes within the blocks was counterbalanced across participants by means of using three different sequences: I–II–III–IV–V, V–III–II–IV–I, and V–III–II–I–IV. The sequences were assigned to participants randomly. We did not use other possible combinatory variants of sequences to keep the accumulated score in all participants above zero throughout the whole experiment duration.

Maintaining a positive balance was needed to prevent participants from experiencing frustration and losing motivation. The accumulated score was converted from points to rubles at 1:1 ratio, and the money was paid to participants after the experiment. On average, participants were paid 420 ± 250 rubles (*M* ± *SD*). The experiment was implemented using the Presentation 14.4 software (Neurobehavioral systems, Inc., Albany, CA).

Pupillometric recording

Pupil size was recorded continuously from the participants' dominant eye using an EyeLink 1000 Plus infrared eye tracker (SR Research Ltd., Canada) at the sampling rate of 1,000 Hz. Pupil size was measured as pupil area in camera pixels using the default eye tracker settings. Before each block, participants completed the EyeLink 1000 9-point calibration procedure.

Data preprocessing

Response time (RT) was measured as an interval from stimulus onset to a button press. Trials with extreme RT values (< 300 ms and > 4,000 ms) were excluded from the analysis; such trials comprised 4.7% of the experimental data. After that, RT data from all valid trials in all blocks jointly were z-transformed within each subject to reduce intersubject variability in RT and make the distribution closer to normal.

Preprocessing of pupillometric data was performed with custom-made scripts in R Studio (Version 3.6.3; R Core Team, 2020) using the "eyelinker" package (Barthelme, 2019). Artifacts related to short interruptions or malfunctioning of pupillometric recording due to blinks or other causes were detected using the following criteria: pupillometric data were missing or the absolute value of the derivative of the pupillometric data exceeded 10 pixels between two adjacent measurements (which were recorded at 1-ms steps). For relatively short artifacts up to 350 ms, which were caused by ordinary eyeblinks, data were replaced by linear data interpolation. If the artifact duration exceeded 350 ms, the respective 5-s epochs (– 2,000 to 3,000 ms relative to the button press) were excluded from the analysis of the pupillometric data.

Next, we downsampled pupillometric data to 20 time points per second by averaging the data into consecutive 50-ms intervals. After that, in order to reduce intersubject variability in pupil size and to make the distribution closer to normal, pupil data were z-transformed within each subject. For this purpose, we used continuous pupillometric data from all five experimental blocks jointly. The baseline for pupillometric measurements was calculated as an average throughout all trials of the experiment (in each participant separately), one and the same for all conditions and timepoints. Then, we converted pupil size at each time point within each epoch into z-scores, i.e., a number of standard deviations from this common reference value. Z-scores computed in this way are similar to a baseline-free approach—yet they provide adjustment for the large interindividual differences in pupil size and make distribution closer to normal (Attard-Johnson et al., 2019).

A common measurement scale for all conditions allowed us to make direct comparisons between conditions and timepoints, with a higher z-score value always reflecting greater pupil size. We used this normalization procedure because the intertrial intervals in the current experiment were rather short compared with the duration of "cognitive" pupillometric responses, which can last up to several seconds (de Gee et al., 2014; Koenig et al., 2018; Lavin et al., 2014; Preuschoff et al., 2011). Pupil dilation caused by an arousing event may thereby change the pretrial baseline for the next trial (the so-called carryover effect). As a result, a conventional baseline correction by means of subtracting the "pretrial baseline" could lead to erroneous estimation of phasic pupil responses (see Attard-Johnson et al., 2019 for a thorough discussion on this matter). Although our normalization procedure did not eliminate the carryover effect, it allowed us to determine the contribution of the carryover effect to the pupil dynamics in adjacent trials. Given that real effects on pupil size emerge slowly, they cannot be expected within the first 220 ms after the stimulus presentation at the trial onset (Mathôt et al., 2018). Thus, carryover effects can be distinguished from effects of the current decision choice by their timing. A relative pupil dilation measured near the time of the stimulus onset (≥ 1,000 ms before the behavioral response), most probably, represents the carryover effect from the previous trial, whereas the relative pupil dilation emerging in close proximity to the response initiation is likely to be a true effect related to the decision making in the current trial.

After that, we excluded from the analysis those epochs during which z-transformed pupil size deviated from zero by more than 3 standard deviations; such trials comprised 0.56% of the experimental data. All pupillometric measurements are reported below as z-scores.

Note, that the trial-related z-scores might be small or negative because the highest positive z-scores were always obtained during the pre-trial baseline characterized by pupil dilation in the absence of any visual stimulation on the screen. Additionally, in order to evaluate slow non-phasic effects, we analyzed the pretrial pupil size (see below). We also supplemented this report by repeating the main analyses using the pupil size data that were baseline-corrected by means of subtracting the pretrial pupil size (see below).

Trial selection criteria and the factors of interest

We assumed that when participants exhibited a stable preference for the advantageous stimulus, they guided their behavior using the utility model that they had acquired through trial-and-error learning during the initial trials of a given block. Therefore, after learning, disadvantageous choices were likely committed against the internal utility model, and thus they could represent intentional directed exploration. Within each block independently, we first identified all trials belonging to such "after learning" periods using the following criteria:

1. Such periods should be preceded by four advantageous choices made in an interrupted succession. The probability of three advantageous choices immediately following one by chance is $(½)^3 = 1/8 = 12.5\%$, which is quite a liberal threshold.
2. The percentage of advantageous choices thereafter until the end of the block should be no less than 65%. One-tailed one-sample binomial test shows that in a sequence of 30 trials (~ number of trials taken into consideration within an experimental block following the first step), the difference between 65 and 50% (random choices) is at a margin of significance at $p = 0.05$.

We used a combination of two rather liberal criteria to distinguish the blocks for which a participant most likely acquired the internal utility model from the blocks where his/her attempts to recognize the advantageous stimulus choice led to failure. We will refer hereafter to such blocks as "after learning" and "no learning" condition, respectively.

Then, we identified trials during which the participants made objectively advantageous choices and disadvantageous choices (i.e., when they selected the stimuli associated with 70% and 30% gain probability, respectively).

Potentially, the disadvantageous choices could be caused solely by negative outcomes of a preceding advantageous choice, which could prompt the participants to immediately change their preference (Gaffan & Davies, 1981; Ivan et al., 2018), i.e., to follow a Win-Stay Lose-Shift strategy (Ellerby & Tunney, 2017). To check whether the disadvantageous choices were mainly caused by a previous loss, we investigated probabilities of transitions leading from the advantageous to the disadvantageous choice. We took the number of all transitions from advantageous stimuli to disadvantageous ones as 100% and calculated the percentage of transitions made after losses. Then, we used a *t*-test to compare this percentage to 100% (as if all transitions to exploration were made strictly as a response to losses) and to 30% (as if transitions to exploration were strictly independent of the preceding feedback).

Next, to investigate the dynamics of behavioral and pupillometric indices over transitions between types of choices made by participants, we used a more detailed trial classification procedure. For this purpose, we took into account not only the choice made during each given trial but also the choices made on the previous one and on the following one. Not all possible combinations could be encountered often enough: only four combinations had a sufficient number of trials during "after learning" condition (> 6 trials for each condition per subject on average), while other possible combinations were rather rare (< 2 trials for each condition per subject on average). Thus, for further analyses, we used the following four levels of "Choice Type" factor (Table 1):

- the "low-payoff" choice (LP) —the disadvantageous choice preceded and followed by advantageous choices;
- the "high-payoff" choice (HP)—the advantageous choice preceded and followed by advantageous choices, thus representing a stable preference for advantageous stimuli;
- the trial preceding the "low-payoff" choice (pre-LP)—the advantageous choice that preceded the disadvantageous and followed the advantageous one;
- the trial following the "low-payoff" choice (post-LP) —the advantageous choice that followed the disadvantageous and preceded the advantageous one.

As mentioned earlier, one could expect that the disadvantageous choices could be provoked by negative outcomes

Table 1 Choice types and overall behavioral statistics

Choice Type	Previous Trial → Current Trial → Next Trial*	No. trials per subject (*M* ± *SD*)		% of Trials per subject (*M* ± *SD*)		Total no. trials		RT (ms)	
		After learning	No learning	After learning	No learning	After learning	No learning	After learning	No learning
HP	A → **A** → A	55.8 ± 43.3	8.4 ± 7.0	60.1 ± 23.8	15 ± 9.7	4965	412	1171 ± 403	1577 ± 717
pre-LP	A → **A** → DA	7.7 ± 6.0	8.5 ± 6.2	11.8 ± 11.7	14.3 ± 6.2	686	416	1393 ± 606	1467 ± 597
LP	A → **DA** → A	6.6 ± 5.3	6.7 ± 5.8	9.7 ± 8	11.1 ± 7.0	587	330	1609 ± 597	1631 ± 703
post-LP	DA → **A** → A	7.2 ± 5.9	8.3 ± 5.6	9.8 ± 6.6	14.9 ± 5.7	639	406	1467 ± 610	1666 ± 669

* A – advantageous choice, DA – disadvantageous choice

of the preceding advantageous choices, thus obscuring the supposedly exploratory nature of disadvantageous choices. In order to check whether this potentially retroactive mechanism did not influence the results substantially, we supplemented this report by repeating the basic analyses related to "Choice Type" factor (see below and supplementary materials) on a smaller subset of trials with two additional restrictions:

1. The outcome of the pre-LP trial within such a sequence of trials should be a gain rather than a loss, and
2. Pre-LP, LP and post-LP trials should constitute uninterrupted sequences of trials in direct succession.

Most importantly for testing the validity of the "directed exploration" hypothesis, we contrasted the same types of choices performed during "after learning" condition vs. choices made in "no learning" condition (Learning factor). In the latter case, the respective trials were selected from blocks during which the participants failed to reach the learning criteria. We suggested that during such blocks, participants did not acquire a proper internal utility model, and any choices made by a participant, regardless its objective advantageousness, represented random rather than directed exploration.

Statistical analysis of RT and pupil size using the linear mixed effects model

We used linear mixed effects models (LMM) at single-trial level rather than repeated measures ANOVA at the grand-average level because LMM method is robust to imbalanced designs across individual cases. Thus, missing data need not result in listwise deletion of cases, and differing numbers of trials per condition are less problematic than in traditional ANOVA (Kliegl et al., 2011). LMM can handle large numbers of repeated measurements per participant, thus making it possible to analyze data from individual trials: this allows accounting for intertrial variability, which would be lost under standard averaging approaches (Tibon & Levy, 2015; Vossen et al., 2011).

Statistical LMM analyses were performed using R software v 4.1.0 (R Core Team, 2021).

We used the following basic procedure, unless specified otherwise. We fitted LMMs on RT and pupillometric data using lme4 package (Bates et al., 2015). We started with the full model, which included relevant fixed effects and their interactions. In the current report, we used the following fixed effects and their interactions: "Choice Type" (4 levels: HP, pre-LP, LP, and post-LP) as described above; "Previous Feedback" (2 levels: gain, and loss—the outcome of the trial that immediately preceded the current one); "Current Feedback" (2 levels: gain, and loss—the outcome of the choice made during the current trial); and "Learning" (2 levels: after learning, and no learning as described above).

All models used for data analysis in the current study included the following random effects intercepts: "Subject" (89 levels), "Block number" (5 levels: one to five—position of a particular block in the sequence of experimental blocks), and "Reinforcement Scheme" (5 levels, the monetary value of gains and losses within a particular experimental block, see above).

The LMMs for the main analyses included the following fixed factors: Choice Type, Previous Feedback, and Learning:

$$\begin{aligned}&\text{Response time} \sim \text{Choice Type} * \text{Previous Feedback} * \text{Learning}\\&+(1|\text{Subject}) + (1|\text{Block number}) + (1|\text{Reinforcement Scheme})\end{aligned} \tag{1}$$

and

$$\begin{aligned}&\text{Pupil size} \sim \text{Choice Type} * \text{Previous Feedback} * \text{Learning}\\&+ (1|\text{Subject}) + (1|\text{Block number}) + (1|\text{Reinforcement Scheme})\end{aligned} \tag{2}$$

Additionally, for illustrative purposes, we aimed to look into the Choice Type x Previous Feedback interaction, within "no learning" and "after learning" conditions analyzed separately. For this purpose, we used the following LMMs for RT and pupil size:

$$\begin{aligned}&\text{Response time} \sim \text{Choice Type} * \text{Previous Feedback} + (1|\text{Subject})\\&+(1|\text{Block number}) + (1|\text{Reinforcement Scheme})\end{aligned} \tag{3}$$

and

$$\begin{aligned}&\text{Pupil size} \sim \text{Choice Type} * \text{Previous Feedback} + (1|\text{Subject})\\&+(1|\text{Block number}) + (1|\text{Reinforcement Scheme})\end{aligned} \tag{4}$$

In the same vein, when Learning factor did not interact with the Choice Type and Current Feedback, we illustrated the Choice Type x Current Feedback interaction, within no learning and after learning conditions analyzed separately. For this purpose, we used the following LMM for the analysis of the pupil size when focusing on the time interval after the feedback onset:

$$\begin{aligned}&\text{Pupil size} \sim \text{Choice Type} * \text{Previous Feedback} * \text{Current Feedback}\\&+(1|\text{Subject}) + (1|\text{Block number}) + (1|\text{Reinforcement Scheme})\end{aligned} \tag{5}$$

For all models, we did a step-down model selection procedure using "step" function implemented in lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017). This procedure performs backward elimination of nonsignificant effects. At the first stage, nonsignificant random effects are eliminated based on the likelihood ratio test (Stuart, Ord, & Arnold, 1999). Then, significance of fixed factors is assessed by using the Kenward-Roger approximation for denominator degrees of freedom (Halekoh & Højsgaard, 2014; Kuznetsova et al., 2017). At each step, the nonsignificant factor with the highest *p*-value is eliminated, and this procedure is repeated until only significant factors remain in the model. Next, we used the simplified model produced by the "step" function. For those factors that remained in the model, we estimated significance using the Satterthwaite approximation for denominator degrees of freedom and obtained type III ANOVA table (package lmerTest, function anova) (Kuznetsova et al., 2017).

Next, to investigate particular contrasts, we performed planned comparisons using the Tukey HSD post-hoc tests (Tukey, 1977) implemented in emmeans package (Lenth, 2021). We did that in two ways. First, we evaluated pairwise differences within the levels of Choice Type factor. Second, to investigate the interference between Choice Type and other factors of interest (Previous Feedback, Current Feedback, or Learning), we evaluated pairwise differences within the levels of the factor of interest split by the levels of Choice Type factor.

Defining the time interval of interest for testing the effects of learning and previous feedback on pupil size

First, we wanted to check whether the pupil size differentiated between advantageous and disadvantageous choices during after learning condition and, if it did, to determine the time span of these choice-specific pupillary responses. The average pupil size measured across the time interval selected this way would allow for testing the role of learning and/or previous feedback in the choice-driven pupil size modulations.

First, we analyzed independently each time point within the epoch ranging from − 1,000 to 2,500 ms relative to time of the response. We ran the following LMM on single-trial data with Choice Type factor (4 levels: HP, pre-LP, LP, post-LP) taken as the fixed effect:

$$\begin{aligned}\text{Pupil size} &\sim \text{Choice Type} + (1|\text{Subject}) \\ &+ (1|\text{Block number}) + (1|\text{Reinforcement Scheme}),\end{aligned} \tag{6}$$

Next, we performed planned comparisons using the Tukey HSD post-hoc test within Choice Type factor (4 levels: HP, pre-LP, LP, post-LP), thus obtaining the statistical significance of pairwise differences between Choice Type factor levels and applied correction for multiple comparisons using the false discovery rate (FDR) method for 70 time intervals at q = 0.05.

As a result, we obtained the time spans of significant differences in pupil size between choice types. Because we were interested in comparing the disadvantageous choices and neighboring trials vs. advantageous trials, we restricted this analysis to three contrasts: pre-LP choices vs. HP choices, LP choices vs. HP choices, and post-LP choices vs. HP choices. For the use in further analyses, we chose the overlap of significant time intervals within these three contrasts. This procedure produced a rather long time interval from − 400 ms to 2,200 ms relative to the behavioral response. Thus, for further analyses, pupil size was averaged within this interval, in each trial separately.

Apparently, this rather long interval was functionally heterogeneous. Thus, additionally, we divided this time interval into three functionally different subintervals, and analyzed them separately (see supplementary materials):

- Decision making and action initiation (− 400 to 0 ms relative to the response),
- Internal outcome evaluation and feedback anticipation (0–1,000 ms relative to the response),
- Matching expected and actual feedback (1,000–2,200 ms relative to the response).

Follow-on analyses of pretrial pupil size and baseline-corrected pupil size using pretrial pupil size

To evaluate slow non-phasic changes in pupil size, that might partly reflect tonic effects, we also analyzed pretrial pupil size, averaged over − 300 to 0 ms relative to the fixation cross onset.

We also repeated the main analyses using the conventionally baseline-corrected data (using the prestimulus baseline) on a trial-to-trial basis; this correction was applied to z-transformed data by means of subtracting pretrial pupil size.

Correlational analysis

In order to test whether commission of disadvantageous choices decreases the profit gained by participants, we calculated Pearson's correlation between the percentage of LP choices and the number of gains, both measures evaluated within "after learning" condition. The percentage of disadvantageous choices was calculated for each participant using the following formula:

$$\text{P\_DA} = \text{N\_DA}/(\text{N\_DA} + \text{N\_A}), \tag{7}$$

where N_DA is the number of all disadvantageous choices, and N_A is the number of advantageous choices for each participant within after learning condition.

In addition, we investigated whether the effects related to disadvantageous choices under after learning condition would be dependent upon how often the participants ventured into such disadvantageous choices. For this purpose, we investigated the relation between the percentage of disadvantageous choices and LP-choice-related changes in both RT and pupil size using Pearson's correlation. LP-choice-related RT slowing was calculated for every subject as the difference between mean RT during LP choices and mean RT during HP choices during after learning condition. LP-choice-related pupil dilation was calculated for every subject as the difference between mean pupil size during LP choices and mean pupil size during HP choices within a post-feedback time interval (1,000–2,200 ms relative to the behavioral response) during after learning condition. In order to reduce statistical noise, we included onto this analysis only those 80 participants who had more than one LP choice.

Since false-positive correlation may result from data unreliability (e.g., outliers are present in the data), we assessed Pearson's correlation significance by way of using permutation tests. This method guarantees a robust and reliable assessment of correlation significance (Higgins, 2004). This algorithm is implemented in perm.cor.test function implemented in package jmuOutlier (Garren, 2019). Briefly, the algorithm of permutation statistics reshuffles the data many times and calculates statistical distribution on simulated data; p-values are assessed as the probability to obtain stronger effects on simulated data compared with real data. We estimated two-sided *p*-values using 20,000 simulations for each analysis.

We plotted scatterplots and respective linear regressions to illustrate the results of the correlational analyses. All statistical analyses were performed using R software v 4.1.0 (R Core Team, 2021).

Results

General behavioral statistics

Before the start of the experiment, participants successfully completed the test for discrimination within pairs of stimuli similar to those used in the experiment. This excludes the possibility that participants could have had any substantial difficulty in perceptual discrimination between the stimuli during the experiment. The overall behavioral statistics is shown in Table 1.

Over the entire period of the experiment (5 blocks), the participants made disadvantageous LP choices on 27.1% ± 14.2% of trials ($M \pm SD$). Eighty-nine of 94 participants fulfilled the learning criteria (4 advantageous HP choices committed consecutively and no less than 65% of advantageous HP choices thereafter until the end of the block) in at least one or greater number of experimental blocks. Because 5 of 94 participants (5.3%) completely failed to learn in any of the experimental blocks, they were excluded from all further analyses. The remaining 89 participants reached learning criteria on 3.7 ± 1.4 blocks ($M \pm SD$) out of 5. Learning criteria were reached by them after 12.4 ± 6.8 trials ($M \pm SD$) out of 40 trials comprising each block.

Participants made disadvantageous choices on 17.0% ± 9.5% of trials within after learning condition, and significantly more often (24.3% ± 22.2%) within no learning condition ($t_{(88)} = 3.00$, $p = 0.004$).

When the stimulus-reward contingency was learned, 42.2% ± 25.3% of all transitions from advantageous to disadvantageous choices were committed after losses. This value is significantly smaller than 100%—the percentage that would have been observed if disadvantageous LP choices were triggered exclusively by losses ($t_{(86)} = -21.28$, $p < 0.001$), and significantly greater than 30%—the percentage of negative outcomes of advantageous choices in the experimental procedure ($t_{(86)} = 4.50$, $p < 0.001$). This means that a previous loss, even if it violated the acquired utility model, did not fully account for the following choice of the disadvantageous option.

Within no learning condition, a very similar pattern was observed: 39.2% ± 19.9% of all transitions from advantageous to disadvantageous choices were committed after losses. The pattern of results was quite similar to that observed after learning: the percentage of transitions after losses was significantly smaller than 100% ($t_{(52)} = -22.2$, $p < 0.001$) and significantly greater than 30% ($t_{(52)} = 3.29$, $p = 0.002$).

The total number of gains negatively correlated with the percentage of LP choices ($r_{(74)} = -0.57$, 95% CI $[-0.71, -0.4]$, $p < 0.001$), thus demonstrating that commission of disadvantageous choices indeed prevented participants from maximizing their cumulative profit, while the optimal strategy for them would be to avoid disadvantageous choices as much as possible (Fig. 2a).

Response time: effects of learning and previous feedback

The LMM for the RT analysis included the following fixed effects: Choice Type (HP, LP, pre-LP, and post-LP), Previous Feedback (gains and losses), Learning (no learning and after learning), and their interactions. The following effects and interactions were statistically significant: Choice Type ($F_{(3,8187)} = 26.2$, $p < 0.001$), Previous Feedback ($F_{(1,8106)} = 6.6$, $p = 0.01$), Learning × Choice Type ($F_{(3,8266)} = 15.5$, $p < 0.001$), Choice Type × Previous Feedback ($F_{(3,8340)} = 5.99$, $p < 0.001$).

Learning effect Planned comparison revealed that Learning × Choice Type interaction was due to the fact that learning affected the RT for the LP and HP choices in opposite directions (Fig. 3a, left panel); it slowed RT during

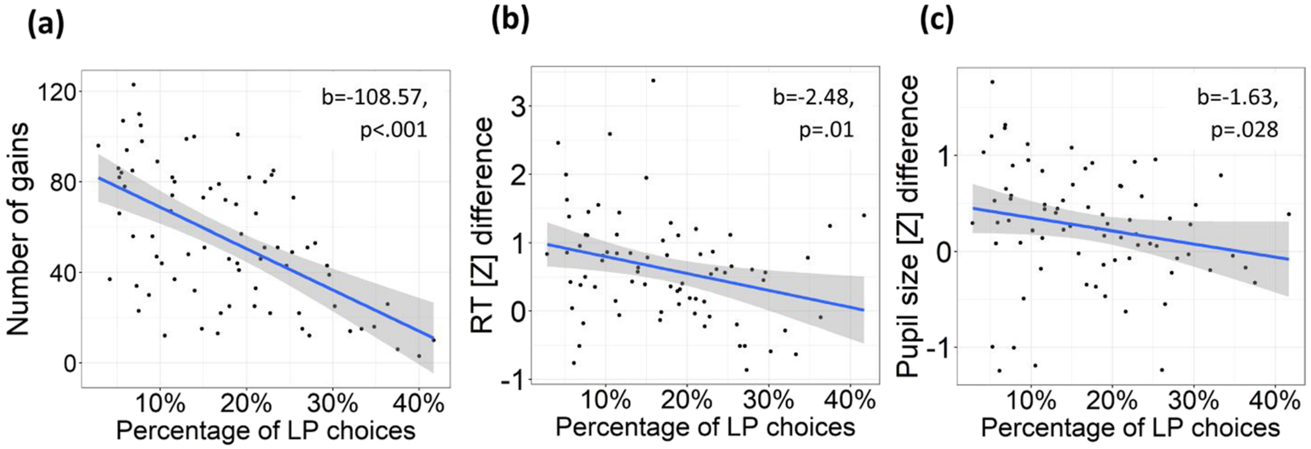

Fig. 2 Scatterplots depicting the relationship under the after learning condition **(a)** between the percentage of LP choices and the number of gains obtained by participants; **(b)** between the percentage of LP choices and RT slowing during LP choices (difference in z-scores between LP and HP choices); **(c)** between the percentage of LP choices and relative pupil dilation during LP choices (difference in z-scores between LP and HP choices, averaged over 1,000–2,200-ms relative to response). Each dot on the scatterplots represents averaged data for one participant. Lines on scatterplots represent respective linear regressions, with shaded areas depicting 95% confidence intervals

LP choices (after learning vs. no learning: Tukey HSD, $p < 0.001$) and accelerated RT during HP choices (after learning vs. no learning: Tukey HSD, $p < 0.001$). Thus, learning of stimulus-reward contingency produced slowing of disadvantageous risky responses and speeding of advantageous HP responses that were committed during periods of stable preference for advantageous stimuli.

Next, we used planned comparisons within the same model to analyze no learning and after learning conditions separately. There were no significant differences in the RT between the LP and HP choice types in no learning condition (Fig. 3a, middle panel). On the other hand, after learning RT became significantly longer for the LP choices compared with all the other choice types, and significantly shorter for HP choices than RT for all the other choice types (Tukey HSD: $p < 0.001$ for pairwise contrasts between LP and HP, and for contrasts between HP/LP with pre-LP, and post-LP choices) (Fig. 3a, right panel). Thus, in after learning condition, but not in no learning condition, decision making regarding disadvantageous choice took more time than that for all types of advantageous choices. Additionally, a moderate yet highly significant response slowing was observed on adjacent trials immediately preceding and immediately following a disadvantageous choice (pre-LP and post-LP choices).

To check whether the effect of RT slowing during disadvantageous LP choices was not a consequence of negative outcomes of a preceding advantageous choice, we repeated the same analysis using a smaller restricted subset of data within uninterrupted sequences of pre-LP → LP → post-LP trials involving only gains on pre-LP trials. The patterns of results concerning the effects of Learning and Choice Type were perfectly preserved in this reduced dataset (supplementary materials, Figure S1). Thus, the effects observed in relation to LP choices, did not result from losses on the preceding trial.

In summary, after learning, which led to formation of the internal utility model, two major changes occurred. First, RT decreased for stable preference for advantageous stimulus, revealing response speeding under a relatively safe strategy. Second, RT slowing was observed for a riskier strategy of disadvantageous choices.

Effect of previous feedback When we considered no learning and after learning conditions together, we observed that losses and gains in the preceding trial differently affected RT for advantageous and disadvantageous choices (Fig. 3b, left panel). RT was significantly slower after losses than after gains, but only in the case of advantageous HP and pre-LP choices (Tukey HSD: p's < 0.001 for loss vs. gain contrasts for both choice types). Thus, we observed post-loss slowing in situations, when both the previous choice and the current choice were advantageous; otherwise, there was no difference in RT between losses and gains.

Because the triple interaction Choice Type × Previous Feedback × Learning was not significant, we could not perform post hoc tests on the data split by all these factors. Instead, for illustrative purposes, we ran a similar LMM on no learning and after learning data subsets separately (Fig. 3b, middle and right panels). Choice Type × Previous

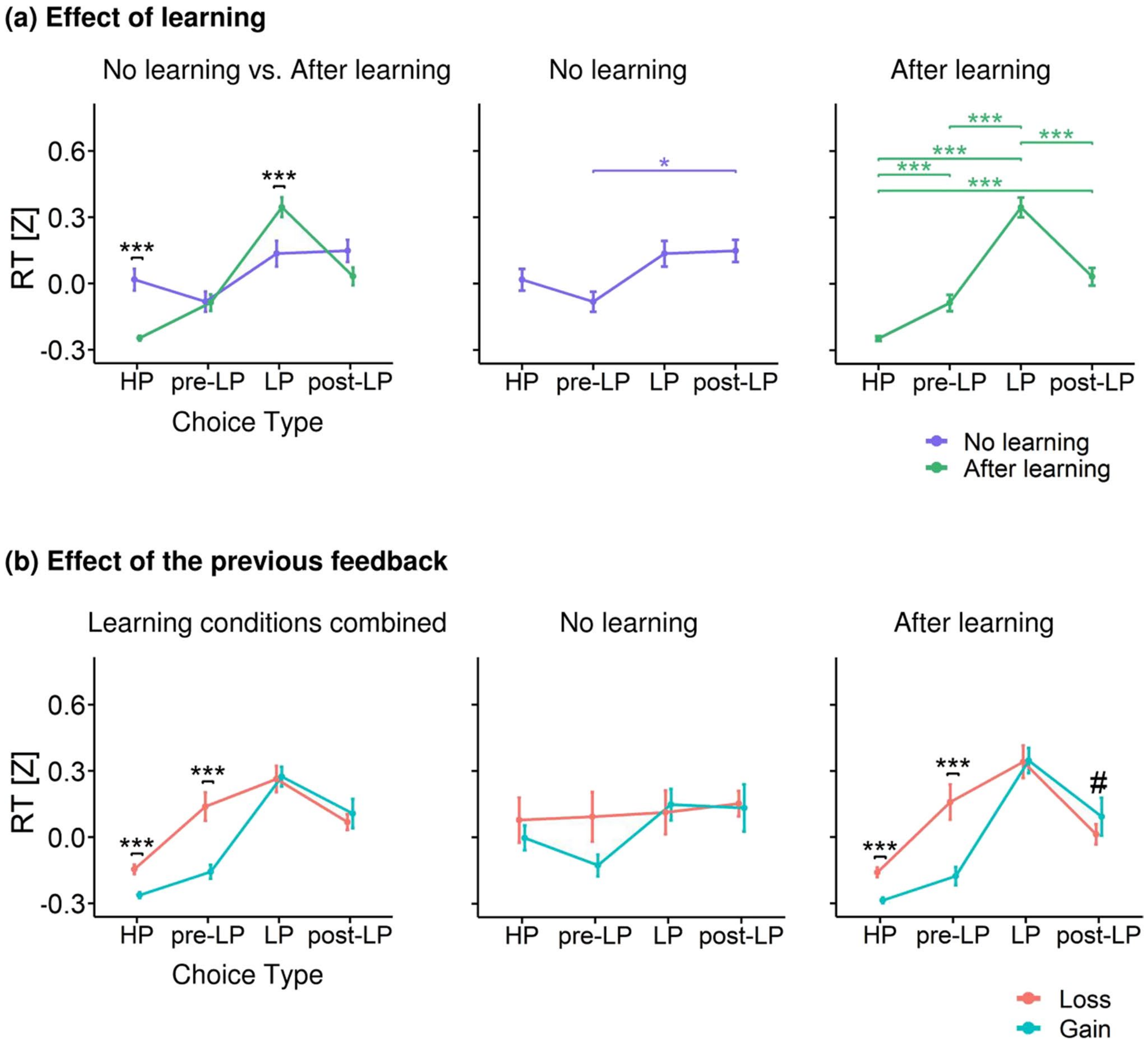

Fig. 3 Response time (z-scored) represented as a function of choice type**. (a)** RT for different choice types under after learning (green) and no learning (slate blue) conditions. Left panel – RT differences between learning conditions within each choice type; middle and right panels – RT differences between choice types within no learning and after learning, respectively. **(b)** RT differences between previous outcomes within each choice type: losses (salmon) vs. gains (turquoise). Left panel – both learning conditions pooled together; middle and right panels – no learning and after learning conditions, respectively. Points and error bars on graphs represent M ± SEM across single trials in all subjects. # $p<0.1$; * $p<0.05$; ** $p<0.01$; *** $p<0.001$

Feedback interaction was significant only within after learning condition ($F_{(3,6846)}=5.77$, $p<0.001$). RT was significantly increased after losses compared with gains within HP choices and pre-LP choices (Tukey HSD: p's <0.001 for both comparisons). No significant differences were found within no learning condition. Thus, the Choice Type × Previous Feedback interaction found in the full dataset was well pronounced mainly under after learning condition, despite a similar trend under no learning one.

Pupil size

Choice-driven modulations of the pupil size timecourses

We were primarily interested to find out whether the type of choice made by participants affected the pupil size. First, we needed to determine the time interval of interest for further analyses, and for this purpose, we ran the following analysis. At the first step, we analyzed each time point independently. We ran the LMMs on single-trial data with Choice Type factor as a fixed effect. Next, we performed planned comparisons using the Tukey HSD post-hoc test for Choice Type factor (4 levels: HP, pre-LP, LP, post-LP). Since at the first step we analyzed each time point independently, at the second step we applied correction for multiple comparisons using the false discovery rate (FDR) method (Benjamini & Yekutieli, 2001) for 70 time points at q = 0.05.

The impact of factor Choice Type on pupil size time courses under after learning condition are represented in Fig. 4. As contrasted with HP choices, the pupil was significantly larger during pre-LP choices (all qs < 0.05 (FDR corrected Tukey HSD) for the time interval from – 400 ms to 2,200 ms relative to the button press; Fig. 4, left panel), LP choices (all qs < 0.05 for the time interval from – 400 ms to 2,200 ms; Fig. 4, middle panel) and post-LP choices (all qs < 0.05 for the time interval from – 1,000 ms to 2,200 ms; Fig. 4, right panel). Thus, for the use in further analyses, we took the pupil size averaged over the overlap of significant time intervals for all contrasts shown above, i.e., from – 400 ms to 2,200 ms relative to the choice response.

It is important to note that for pre-LP and LP choices the effects started around 400 ms before the button-press, while they were absent at earlier times (Fig. 4, left and middle panels). Thus, these pupil dilations apparently were not just carryover effects inherited from the previous trials (Mathôt et al., 2018).

Effects of learning and previous feedback

The model for this analysis included the following fixed effects: Choice Type, Previous Feedback, Learning, and interactions between them. Factors Choice Type ($F_{(3,8385)} = 7.91$, $p < 0.001$) and Previous Feedback ($F_{(1,8422)} = 16.27$, $p < 0.001$) were significant (note that significance for Choice Type factor was partially related to the method that we used to define the time interval for the analysis). More importantly, there were statistically significant interactions: Learning × Choice Type ($F_{(3,8379)} = 11.4$, $p < 0.001$) and Choice Type × Previous Feedback interaction ($F_{(3,8403)} = 5.31$, $p = 0.001$).

Learning Effect Planned comparison using the Tukey HSD post-hoc test within Learning factor split by Choice Type revealed that pupil size was significantly increased for LP choices in after learning condition compared with no learning condition (Tukey HSD, $p = 0.04$) (Fig. 5a, left panel). On the contrary, pupil size was significantly decreased for HP choices in after learning condition (Tukey HSD, $p < 0.001$). Thus, for pupil size we observed the same pattern of the learning effects as that for RT.

Then we probed how Choice Type influenced pupil size within no learning and after learning conditions analyzed separately. In contrast to after learning, in the no learning condition, there were no significant differences between choice types (Fig. 5a, middle panel). Thus, in the absence

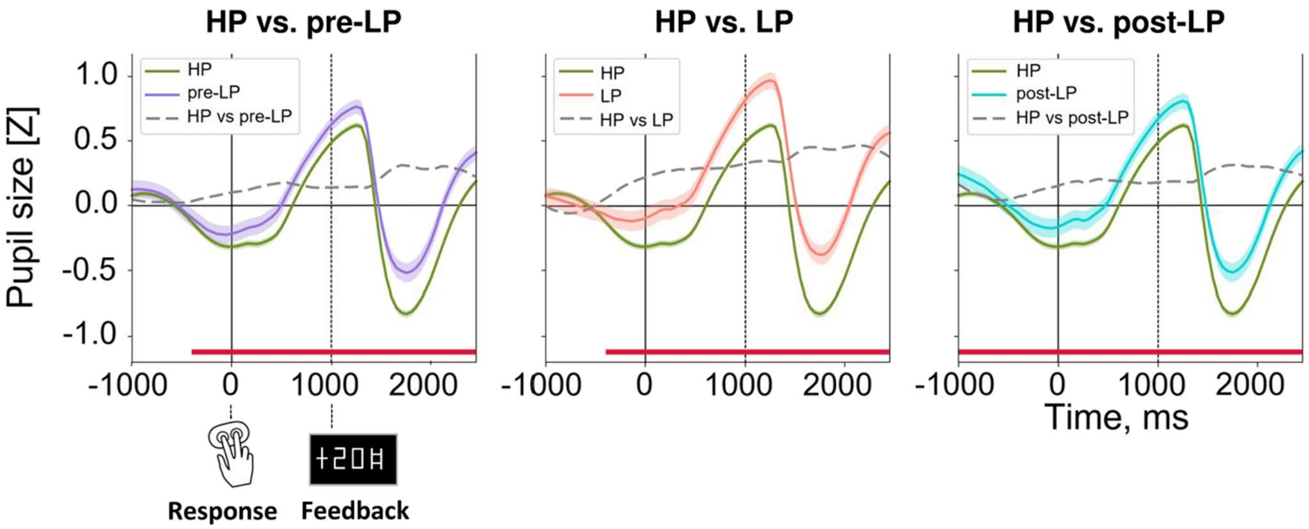

Fig. 4 Time courses of pupil size (z-scored) during choice and after feedback in the after learning condition**.** From left to right: The pre-LP (violet), LP (red), and post-LP (marine blue) choices in comparison with the HP choice (green). The dashed curve in each graph corresponds to the time course of the difference in the pupil size between the respective choice type and the HP choice. Solid and dashed vertical lines correspond to button press (zero point) and feedback onset respectively. Curves and shaded areas represent *M* ± *SEM* across single trials in all subjects. Dark red lines at the bottom of each graph indicate significant differences between the respective choice type and HP choice ($p < 0.05$, FDR corrected)

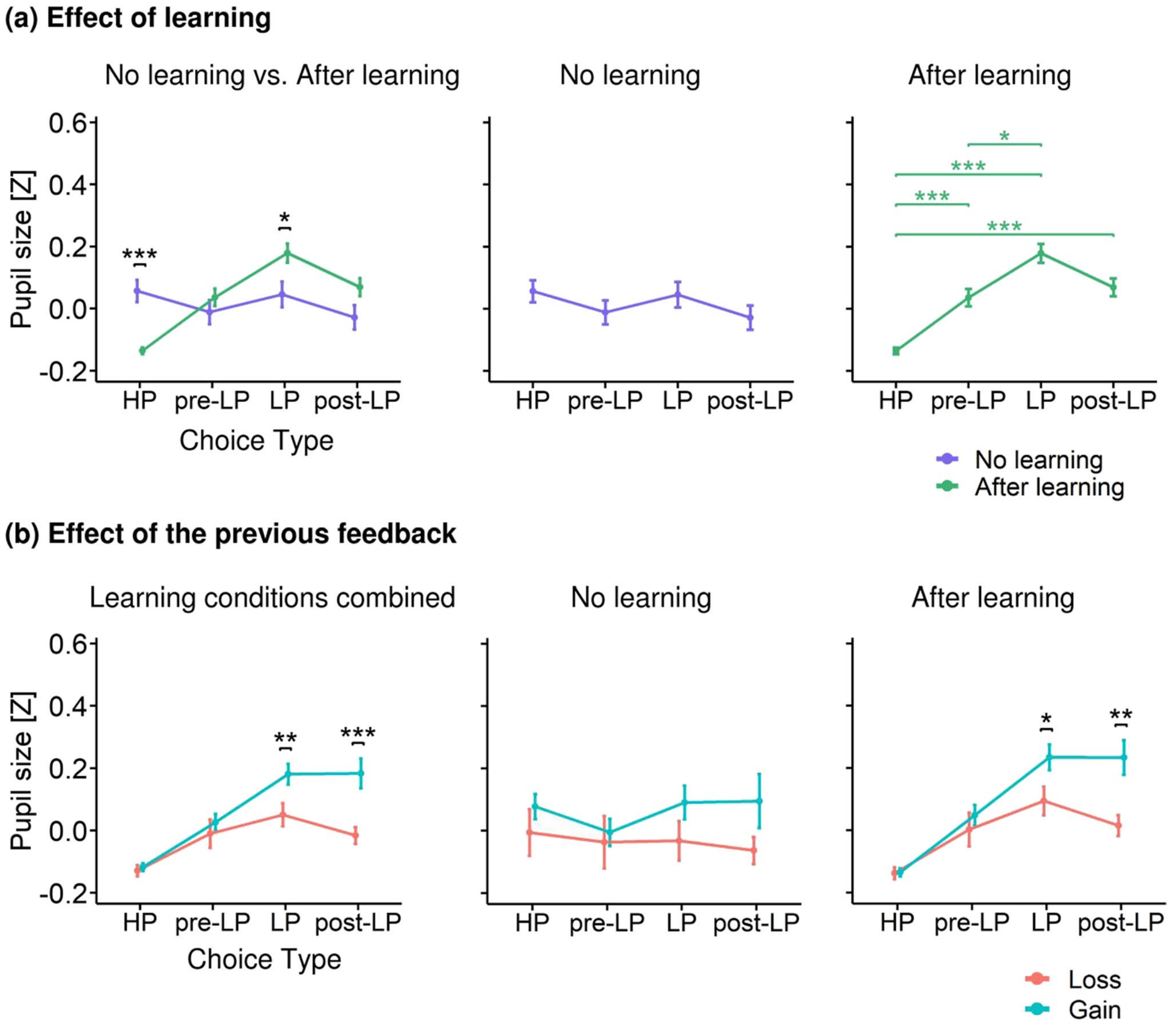

Fig. 5 Pupil size (z-scored) represented as a Function of Choice Type. **(a)** Pupil size averaged within the interval – 400 to 2,200 ms relative to the response (button press) for different choice types under the after learning (green) and no learning (slate blue) conditions. Left panel – pupil size differences between learning conditions within each choice type; middle and right panels – pupil size differences between choice types within no learning and after learning, respectively. **(b)** Pupil size differences between previous outcomes within each choice type: losses (salmon) vs. gains (turquoise). Left panel – both learning conditions pooled together; middle and right panels – no learning and after learning conditions, respectively. All other designations as in Fig. 3

of the internal utility model, pupil size did not depend upon the type of choice made by participants.

In the after learning condition (Fig. 5a, right panel), pupil size was significantly greater during LP choices compared with HP choices (Tukey HSD, $p < 0.001$) and pre-LP choices (Tukey HSD, $p = 0.011$). Additionally, pupil size was significantly greater during pre-LP and post-LP choices compared with HP choices (Tukey HSD, $p < 0.001$). Thus, in convergence with the RT data, pupil size after learning was increased during disadvantageous choices compared to advantageous choices; it also was moderately yet significantly increased during advantageous choices on adjacent trials that immediately preceded and immediately followed disadvantageous choices (pre-LP and post-LP).

At least some of the disadvantageous choices were committed after losses and could be immediately caused by them. In order to exclude the possibility that this potentially retroactive mechanism could influence the results, we repeated the same analysis within the Choice Type factor using a smaller restricted subset of data with uninterrupted sequences of pre-LP → LP → post-LP trials involving only gains on pre-LP trials. Comparisons in this reduced dataset closely reproduced the patterns of effects described above for a full dataset (supplementary materials, Figure S2). Thus, the effect of pupil dilation during disadvantageous choices was not caused by losses on trials preceding them.

In summary, we found that after learning, which led to formation of the utility model, two major changes occurred. First, pupil size became smaller for the repetitive HP choices, during which participants exhibited a stable preference for the advantageous stimulus, i.e., were keeping with a relatively safe strategy. Second, pupil size was increased for unsafe disadvantageous choices.

Effect of previous feedback As with RT, the sign of the previous feedback differently affected pupil size depending on the Choice Type, as reflected by significant interaction Choice Type × Previous Feedback, but the pattern was qualitatively different (Fig. 5b, left panel). In contrast to RT, pupil size was greater after gains than after losses, but only during LP and post-LP choices (Tukey HSD, $p = 0.023$ and $p = 0.002$, respectively); both choice types implied a switch from one strategy to another (from exploration to exploitation in post-LP choice and from exploitation to exploration in LP choice). No significant differences in pupil size between gains and losses were observed for HP and pre-LP choices.

Again, to analyze Previous Feedback influence on pupil size within no learning and after learning conditions separately, for illustrative purposes, we ran a similar LMM on no learning and after learning data subsets separately (Fig. 5b, middle and right panels). The analysis revealed Choice Type × Previous Feedback significant interaction within after learning only ($F_{(3,6823)} = 3.94$, $p = 0.008$): pupil size was greater after gains compared with losses for LP and post-LP choices (Tukey HSD, $p = 0.023$ and $p = 0.002$, respectively), but not in HP and pre-LP choices.

Subintervals within the choice-related time period

In the main analyses of the pupil size, we used a rather long choice-related time interval, which may be functionally heterogeneous. Therefore, we divided this time interval into three successive functional subintervals: prior to a choice (decision making and action initiation), between a choice response and feedback signal (internal outcome evaluation in anticipation of the feedback), and after feedback (matching expected and actual outcome). Then, we analyzed each subinterval independently using the same statistical procedure as that used for the full choice-related time interval (supplementary materials, Figures S3 and S4). In all three subintervals, we observed the pattern of results highly compatible with that obtained in the main analysis, with the latest subinterval (1,000–2,200 ms relative to the response) manifesting the most pronounced statistical effects. Importantly, in each of these subintervals, pupil size was greatest during disadvantageous trials, with a similar yet attenuated effect for advantageous choices on adjacent trials immediately preceding and immediately following disadvantageous trials compared with the stable preference for advantageous choices (supplementary materials, Figure S3). Interaction Choice Type x Previous Feedback was significant in the second and third subintervals. Again, the pattern of results on these subintervals was similar to that obtained on a full response-related time interval of interest (supplementary materials, Figure S4). This suggested that a relatively increased pupil size reflected a protracted common process that affected different stages of decision making in regard to LP choices made during the after learning condition.

Pretrial pupil size

To evaluate whether the learning-induced effects found for choice-related pupil modulations involved slow tonic components and/or carryover effects from previous trials, we additionally analyzed the pretrial time interval (− 300 to 0 ms relative to fixation cross onset) using the same statistical procedure as that used for choice-related interval (Fig. 6a). The following effects were statistically significant: Choice Type ($F_{(3,7539)} = 10.98$, $p < 0.001$), Learning ($F_{(1,4504)} = 9.64$, $p = 0.002$), Learning × Choice Type ($F_{(3,7535)} = 6.84$, $p < 0.001$). Planned comparisons of Learning × Choice Type interaction revealed that learning success mainly increased pre-trial pupil size for post-LP choice (after learning vs. no learning for the post-LP: Tukey HSD, $p < 0.001$) but not for the LP choice itself (Fig. 6a, left panel).

In the no learning condition, pre-trial pupil size did not discriminate between choice types (Fig. 6a, middle panel). In the after learning condition, the main choice-related distinction was a small “baseline” pupil size for HP choices compared with pre-LP, LP, and post-LP choices (Tukey HSD, $p = 0.003$, $p < 0.001$, and $p < 0.001$, respectively) (Fig. 6a, right panel). Yet, the greatest pre-trial pupil size was observed not before the LP choice itself but before post-LP choice (pre-LP vs. LP after learning: Tukey HSD, $p < 0.001$).

Thus, unlike pupil size during a choice, its modulations in the pre-trial period did not accentuate the impact of learning on pupil size during the disadvantageous choice. The dramatic increase in the “baseline” pupil size preceding post-LP choices is likely a carryover effect lasting from the choice-related pupil dilation in the previous LP trial (compare Fig. 5a, right panel and Fig. 6a, right panel). Yet, the fact that during the after learning condition the pupil was greater before pre-LP and LP trials compared with HP trials hints at some tonic effect, which contributed to disadvantageous choices.

Previous feedback did not affect pretrial pupil size in both learning conditions. Thus, the impact of previous feedback on the pupil size was mainly related to the choice response itself.

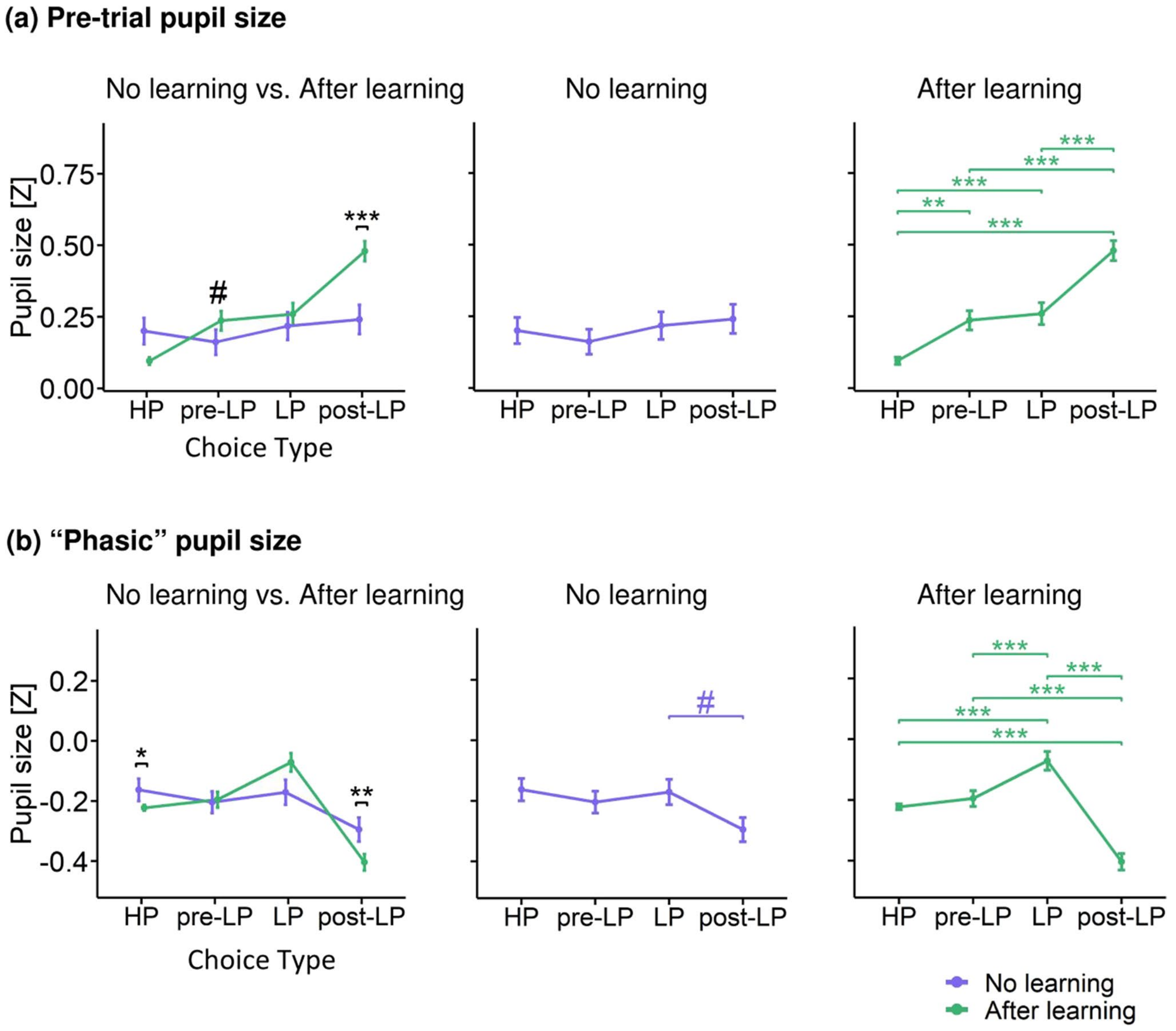

Fig. 6 Baseline and phasic changes in pupil size (z-scored). **(a)** Pre-trial pupil size averaged within the interval – 300–0 ms relative to the fixation cross onset for different choice types under after learning (green) and no learning (slate blue) conditions. **(b)** Phasic pupil size† for different choice types under after learning (green) and no learning (slate blue) conditions. Left panel – pupil size differences between learning conditions within each choice type; middle and right panels – pupil size differences between choice types within no learning and after learning, respectively. All other designations as in Fig. 3. † Note: Phasic pupil was calculated as the difference between the z-scored pupil size averaged within the interval – 400–2,200 ms relative to the response onset and pretrial pupil size. Pupil size was z-scored within each participant using all time points across all trials and pretrial intervals. Consequently, subtraction between two values (pretrial and trial) produced negative z-scores for the phasic changes in the pupil size because of the pupillary light reflex evoked by luminance increment during the stimulus presentation

Baseline-corrected pupil size

As an additional step, to corroborate the phasic nature of the pupil effects described in the main analysis, we used the conventional baseline-corrected choice-related pupillary response magnitude, and applied the same statistical model as in the main analysis. The following factors were significant: Choice Type ($F_{(3,7529)} = 16.87$, $p < 0.001$), Previous Feedback ($F_{(1,7525)} = 8.69$, $p = 0.003$), Learning × Choice Type interaction ($F_{(3,7522)} = 3.97$, $p = 0.008$).

Baseline correction preserved the basic pattern of results for the learning effect on choice-related pupil modulations except for the post-LP choices (Fig. 6b). Learning × Choice Type interaction was partially due to the opposite direction of learning-induced pupil changes for the HP and LP choices, similar to that detected in the main analysis (Fig. 6b, left panel). In addition, under the after learning condition, the maximal baseline-corrected pupil size was observed for the LP choices compared with all other choice types, which also is in concordance with the previous finding (compare Fig. 6b, right panel with Fig. 5a, right panel. Concurrently, baseline—corrected pupil size became minimal for post-LP choices—although apparently, this is a technical result of baseline subtraction. Inspection of Fig. 6a, right panel, shows that the pretrial pupil size for post-LP choices was enormously high, most likely due to the long-lasting impact of disadvantageous LP choice on the pupil size that sustained through the whole intertrial period (compare Fig. 6b, right panel with Fig. 6a, right panel). The occurrence of profound technical interactions with the pretrial baseline requires caution when using unsupervised usage of baseline correction for studying a "phasic" pupil response, especially when the intertrial interval is relatively short.

Thus, as could be expected from the main analysis and the analysis of the pretrial pupil size, the prestimulus baseline-correction preserved the basic effects of learning and choice type on pupil dilation for exploratory and preexploratory choices but eliminated the ones for postexploratory choices. In other words, "prestimulus baseline-free" and "prestimulus baseline-correction" approaches yielded near-identical outcomes, leading to similar conclusions. These concordant results strongly suggest that pupil dilation during exploratory and preexploratory choices involves a strong phasic component. However, some systematic carryover effects were observed for postexploratory choices, whereby pupil dilation from an arousing exploratory choice still influenced pupil size on the next postexploratory trial.

To sum up, the additional analysis confirmed that the increased pupil size during disadvantageous choices made after learning involved a strong phasic component.

Correlational analysis

Within the after learning condition, there was a significant negative correlation between the percentage of LP choices and LP-choice-related RT slowing ($r_{(73)} = -0.29$, 95% CI $[-0.49, -0.07]$, $p = 0.01$) (Fig. 2b). In other words, the more often subjects made LP choices, the less RT slowed down in those choices compared with HP choices. Also, there was a similar correlation for pupil size ($r_{(73)} = -0.25$, 95% CI $[-0.45, -0.028]$, $p = 0.03$) within the 1,000–2,200-ms time interval after response onset (Fig. 2c); pupil dilation during LP choices was diminished in those participants who committed LP choices more often.

Discussion

When offered a choice between two alternatives in a standard probability learning task, people occasionally shift their preference toward the option yielding a lesser (mathematical) expectation of the reward. To explain this suboptimal behavior, it has been hypothesized that people intentionally seek patterns in the sequences of outcomes, even though none are present and the probabilities of reward are held constant (Ellerby & Tunney, 2017; Unturbe & Corominas, 2007). From this perspective, rare transitions from objectively advantageous to disadvantageous choices may represent a directed exploration strategy that guides the choice toward an option with uncertain payoff (Wilson et al., 2021).

To test this hypothesis, we contrasted pupil size and response time for LP and HP choices before and after the participants became aware of response-reward contingencies. We found that LP choices were linked to a profound RT slowing and greater pupil dilation but only if the internal utility model has already been acquired by a participant. For the pupil size, this effect was strongly amplified for the LP choice, which immediately followed the gain as compared with loss in the preceding choice. Critically for the hypothesis tested, LP versus HP difference in pupil size involved a strong phasic component and was temporally related to the behavioral choice, but neither to the stimulus itself, nor to the feedback about the choice outcome.

Importantly, our analysis of probabilities of transitions from advantageous to disadvantageous choices proved that the disadvantageous choices themselves were not simply caused by negative outcomes of a preceding advantageous choice, i.e., most disadvantageous choices did not result from a simple Win-Stay Lose-Shift strategy (Ellerby & Tunney, 2017; Gaffan & Davies, 1981; Ivan et al., 2018).

In the following discussion, we will argue that the pattern of results obtained suggests that the rare, objectively disadvantageous choices, that violate the acquired internal utility model, do represent self-generated exploratory behavior. This exploratory strategy causes shifts in choice priorities in favor of information seeking, while its autonomic and behavioral concomitants are mainly driven by a conflict between the behavioral plan of the intended exploratory choice and its predominant alternative, which has already proven to be more rewarding during previous trials.

First, we wanted to ascertain that the observed RT and pupil size dynamics cannot be explained by general cognitive processes, nonspecifically related to the value-based decision formation – sustained attention, internal error detection, outcome monitoring related to the external feedback, and reaction to a previous loss (for review see Zenon, 2019).

The simplest explanation of LP choices is the loss of attention to stimulus display during objectively disadvantageous choice. Response slowing was previously observed during continuous attentional tasks on the trials during which, according to participants' reports, his/her attention was disengaged from the current task either due to the involvement in the inner thoughts or simply due to the decrement of alertness (Cohen & van Gaal, 2013; Dyson & Quinlan, 2003; O'Connell et al., 2009; Ratcliff & McKoon, 2008). However, in sharp contrast to the LP choices in our experiment, response slowing during unfocused attentional states was associated with reduced (not increased as during LP choices) task-evoked phasic pupil dilation (Figs. 3a and 5a) (Unsworth & Robison, 2016). Convergent, instead of divergent, changes in the phasic pupil dilation and the response time during LP choices refute the suggestion that they originated from attentional lapses.

Still, another possibility is that response slowing and pupil dilation during LP choices are driven by internal detection of accidentally committed erroneous response. In the experimental tasks requiring participants to learn arbitrary association between visual stimuli and specific response, RT slowing accompanied with phasic pupil dilation is commonly observed not only after error commission (post-error slowing) (Critchley et al., 2005; Wessel et al., 2011), but also during the erroneous response itself (error slowing); such effect was tentatively ascribed to an error-evoked orienting response (Murphy et al., 2016). At first glance, "error-evoked" explanation seems plausible here, because relatively greater pupil dilation in the LP vs. HP trials emerges as a result of a conscious appraisal of LP choices as disadvantageous (the effect was present exclusively in the after learning condition, Figs. 3 and 5), and hence "erroneous." This "error-detection" explanation, however, is difficult to reconcile with our finding of highly significant pupil dilation and RT slowing in the prelude to an LP choice—a pre-LP trial (Figs. 3 and 5), when no erroneous response was committed, and a participant undertook a "correct" advantageous choice. Also, in contrast to our data, preerror speeding rather that slowing is commonly observed (Dudschig & Jentzsch, 2009), while we observed slower responses on pre-LP trials compared with HP trials.

The third cognitive process putatively involved in the LP choices could be external outcome monitoring; it implies that phasic pupil dilation is caused by the negative feedback contingent with the objectively disadvantageous choice (Satterthwaite et al., 2007). Refuting this possibility, pupillary response during such choices emerges and sustains throughout almost the whole decision-making interval, long before the feedback was provided (Fig. 4).

Apart from the cognitive processes involved in the LP choices themselves, a participant's reaction to the negative outcome of the previous "correct" choice also should be considered as a possibility. Post-loss slowing has been reported for some gambling tasks (Brevers et al., 2015; Goudriaan et al., 2005), although pupil measurements in these studies were lacking. In order to check whether the previous loss substantially influenced our data, we repeated the basic analysis on the RT and pupil size data using uninterrupted sequences of trials, in which the LP choices were committed exclusively after wins, i.e., after those HP choices that were rewarded (supplementary materials, Figures S1 and S2). In such a restricted dataset, the RT and pupil size effects remained highly significant, thus dismissing the predominant role of sensitivity to immediate previous loss in a subject's decision to switch to the obviously disadvantageous choice.

After refuting alternative explanations of our RT and pupil findings, we argue that a participant's decision to seek new alternatives (directed exploration) seems to be the most plausible explanation of behavioral and pupil changes evoked by spontaneous LP choices. This account is based on the findings in the literature that relate a slow RT and choice-evoked pupillary response to information processing and updating the internal model in the brain (see Zenon, 2019 for review). A critical distinction between the current and previous pupil studies of exploration/exploitation dilemma is the nature of exploratory choice itself. The previous pupillometric studies investigated so-called random exploration (Wilson et al., 2021); in these studies, exploratory choices were explicitly encouraged by a gradual decrease in the reward probability for a preferred choice in a restless multiarmed bandit task (Gilzenrat et al., 2010; Jepma & Nieuwenhuis, 2011). Our experimental design was fundamentally different, because it did not involve any systematic changes in the utility of a particular choice and the decider had known the probability of likely outcomes from the previous experience. Thus, our findings provide the first evidence for pupillary response accompanying self-generated exploratory decisions such that participants intentionally choose a risky exploratory option against their behavioral bias toward value-driven choices.

Although the differentiation between random (purely uncertainty-driven) and self-generated or directed exploration (intentional information seeking) is a long-standing problem in psychological literature (Berlyne, 1966; Wilson et al., 2021), physiological concomitants of the directed exploration are largely unknown (but see Zajkowski et al., 2017). In this respect, our findings add an important new dimension to the existent knowledge about the relation of pupil dilation as a measure of LC-NA arousal to human exploratory behavior. Moreover, inclusion of the no learning and the after learning conditions in the present study allowed us to examine the change in choice-related pupillary response from random to self-generated exploration.

The question that we were seeking to answer was whether the concept of subjective uncertainty/surprise used to explain pupil-related LC-NA arousal during exploratory choices (Van Slooten et al., 2018) also can be applied to self-generated exploration or whether different decision processes are invoked depending on the source of uncertainty.

One possibility is that the effect of uncertainty played a similar role in producing transient pupillary responses during both random and self-generated exploratory choices. Hypothetically, in the no learning condition, during which the reward structure remained largely unknown for participants, either choice was "random" and was characterized by an equal uncertainty in the prior belief regarding the outcome. This can explain why a pupillary response to either choice did not distinguish the LP and HP choices in the no learning condition (Fig. 5a, middle panel). As soon as the participants learned to prefer choices that had been probabilistically associated with positive outcomes (after learning condition), the uncertainty was greatly reduced for such advantageous choices, leading to a highly significant attenuation of both choice-related pupillary response and response time costs (Figs. 3a and 5a). Still, ambiguity remained whether other response strategies incorporating occasional risks—choices with a low payoff probability—might produce better total outcomes than the status quo. On the basis of the empirical consensus of association between pupillary response and subjective uncertainty (Richer & Beatty, 1987; Satterthwaite et al., 2007; Urai et al., 2017; Van Slooten et al., 2018), one might predict that the choice-related increase in pupil size, although attenuated for safe choices after learning, would be preserved for the risky explorative ones. This learning-related difference in pupillary responses between the safe and risky choices was exactly what we observed in our data (Fig. 5a).

Thus, at first glance, our findings match well with the principle derived from computational modelling (Jepma & Nieuwenhuis, 2011; Urai et al., 2017; Van Slooten et al., 2018)—phasic pupil dilation is proportional to the subjective estimate of uncertainty. On the other hand, in our experimental settings, the implicit conflict between "safe" and "risky" explorative options is an inevitable consequence of self-generated exploratory choice. The learned value of the objectively advantageous choice is known to create an unconscious value-driven bias (for review see Anderson, 2016) that can interfere with the effects of voluntary endogenous selection determined by the goal of a subject (Preciado et al., 2017). Implicit conflict with this unconscious bias arises when a preferable "safe" action plan is overruled by a deliberate "risky" exploratory decision. Previously, the phasic increase in pupil size was found to be a robust measure of implicit conflict between task-appropriate and habitual automatic responses in a color-naming Stroop task (Laeng et al., 2011). In the context of a value-driven choice, pupil dilation was mainly studied under an explicit conflict, which was parametrically manipulated by changing the already learned differences in the likelihood of reward between two alternatives. Specifically, phasic pupil dilation was found to closely track a degree of explicit conflict, being inversely proportional to the difference in the reward probability for each of the alternative options (Van Slooten et al., 2018). A similar effect was described for the intertemporal choice paradigm when the degree of conflict between the competing subjective preferences for immediate or delayed reward was also parametrically manipulated and formally modelled (Lin et al., 2018). Notably, strongest dependency of pupillary response and decision time cost on the degree of explicit conflict was found for appetitive conditions, i.e., a choice between two conflicting equally desirable win–win action plans (Cavanagh et al., 2014). Thus, when each alternative has significant advantages and disadvantages, people often experience conflict that makes the choice aversive and causes choice-related pupil dilation.

A strong influence of explicit conflict between the two action plans on pupillary response and RT supports our hypothesis that implicit conflict pertinent to self-generated exploratory choices has a similar effect on pupil dilation. In the latter case, conflict arises from competition for action selection between the unconsciously biasing effect of previously rewarded action and voluntary decision, which is shifting choice priorities in favor of information seeking.

We therefore tried to distinguish between "subjective uncertainty" and "conflict" effects by analyzing the pupil dilation in the "safe" choices that immediately preceded and followed the deliberate "risky" choice, as well as by considering the effects of the previous feedback sign on pupillary response. First, both behavioral and pupil results suggest a protracted decision formation such that an intentional exploratory decision was actually made during the preceding trial, maintained, and then enacted during exploratory choice itself (Figs. 3a and 5a). This finding is fully compatible with the previous reports demonstrating that pupillary response associated with the internal state of random exploration begin to develop on the trial preceding the exploratory choice itself (Gilzenrat et al., 2010; Jepma & Nieuwenhuis, 2011; Jepma et al., 2010). However, the pure uncertainty account is difficult to reconcile with our finding that phasic pupil response and response time remained relatively high on the trial immediately following self-generated exploratory choice, when a participant returned to the "safe" choice strategy with a knowingly high payoff probability (Figs. 3a and 5a). Neither effect of uncertainty alone can explain why self-generated exploratory choice (during after learning condition) elicited greater pupil dilation and slower response time than a random exploratory choice with completely unpredictable choice outcome (during no learning condition) (Figs. 3a and 5a). This rather suggests a cumulative contribution of uncertainty regarding the desired outcome and a conflict with a value-driven bias in the self-generated exploration.

Further evidence for the contribution of conflict dimension to the self-generated exploratory choice is provided by the amplifying role of the previously obtained positive feedback on the increased pupillary response. The previous reward as compared to punishment was associated with a greater pupil dilation in both “risky” explorative choices and “safe” post-explorative choices, while this effect was absent for the two other “safe” advantageous choice types (HP-choice and pre-LP choice) (Fig. 5b). Importantly, in both choice types sensitive to the previous reward, the participants shifted to a response that was incongruent with the positive outcome of the previous choice. They changed their action plan toward exploration after being rewarded for the exploitative action (LP-choice), or vice versa, returned to exploitative strategy despite the successful outcome on the previous, explorative, action (post-LP choices). Because the other “safe” choices followed uninterrupted history of the previous frequently rewarded choices, they did not conflict with a previously rewarded action (supplementary materials, Figure S2). This finding indicates that the pupil dilation during self-initiated exploration is likely to reflect more than one process occurring concomitantly.

Notably, for RT measurements, the effect of the previous reward during explorative choices and postexplorative choices was not observed (Fig. 3b). This may be explained by the powerful effect of the preceding loss on response time—i.e., post-loss slowing (Brevers et al., 2015; Goudriaan et al., 2005)—but not on pupil size (Fig. 5b) that was seen in both the HP- and pre-LP choices. This generally adaptive tendency, which serves to promote caution in decision making after losses, may counteract the opposite effect of previous reward on response speed in LP and post-LP trials.

To sum up, while subjective uncertainty is likely to play an important role in phasic pupil dilation caused by directed exploration, it is hardly the only factor determining strong enhancement of pupil size observed here in such choices. Pupil dilation is rather caused by the combined effect of subjective uncertainty regarding exploratory choice outcome and the conflict signal broadcasting that the intended exploratory action violates the internal utility model, which favors the frequently rewarded alternative.

The self-generated decisional challenge whether to explore a set of alternative choices or stick to the opportunity to make a “default” choice suggests a comparison process taking place within the anterior cingulate cortex (ACC). Given that ACC in generally involved in the comparison between the outcome values of different choice options (Kolling et al., 2016), and specifically in conflict monitoring processes (Botvinick et al., 2004; Shenhav et al., 2013), the detection of conflict with the inner utility model in self-initiated exploratory choices may drive transient changes in LC-NA-mediated arousal, which, in turn, increases phasic pupil dilation observed here.

This speculation is consistent with the recent monkey study directly demonstrating that in some cases, pupil related modulations of spontaneous neuronal activity reflect signals occurring first in ACC and then being transmitted to the LC and other subcortical and cortical structures (Joshi et al., 2016). It also is in accord with the human findings showing that the phasic pupil dilation during explicit high-conflict appetitive choices correlates with increased mediofrontal activation (Cavanagh et al., 2011, 2014). Therefore, phasic pupil dilation triggered by the implicit conflict in our experiment may reflect downstream signal of conflict processing in ACC.

Evidence from multiple psychological and physiological research indicate that conflict is emotive and triggers a negatively valenced affective state accompanied by changes in heart rate, skin conductance, body temperature, pupil response, and muscle tone (for review see Saunders et al., 2017). One cannot exclude, therefore, that the negative emotion triggered by implicit conflict may contribute to pupil dilation that we observed for self-generated exploratory choices. By changing the inner affective state, learning from the history of rewards and punishments may reach perception of knowing without conscious awareness (Bechara et al., 1997)—the process of nonconscious information processing that has been referred to as System 1 (Kahneman, 2003). Functionally, changes in the inner physiological state may serve as a subconscious warning signal informing a decision maker that his/her deliberate action plan violates the inner brain model for utility of the intended action.

Conclusions

The study demonstrates that response slowing and augmented pupil-related phasic arousal characterize a participant’s decision to take risk of information seeking by choosing an uncertain alternative over the rewarding one in a two-choice probabilistic learning task. The behavioral and pupillometric findings also suggest that such directed exploration is bound to a conflict between the deliberate explorative choice with uncertain outcome and the inner bias to select the option with the highest value. The close relationship between self-generated choices and pupil-related arousal makes a simple probabilistic learning task a complementary instrument for studying neural underpinnings of directed exploration and the underlying pathophysiology of its abnormalities in mental disorders.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13415-022-00996-z.

Acknowledgements We are grateful to Ivan S. Pozdniakov for invaluable help in recruiting participants. The study was implemented at the unique research facility 'Center for Neurocognitive Research (MEG-Center)'.

Authors' contributions GK, AP, TS, and BC contributed to conception and design of the study. GK, KS, AP, and AR were responsible for the experiment and data acquisition. GK, KS, AP, and VM performed statistical analysis of the data. KS, VM, AR, TS, and BC prepared illustrations. GK, KS, VM, TS, and BC interpreted the data. GK and BC wrote the first draft of the manuscript. GK, KS, VM, AR, TS, and BC wrote sections of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version.

Funding This study was supported by Russian Science Foundation (project # 20–18-00252).

Data availability Data are openly available at 10.6084/m9.figshare.16825570. The experiment was not preregistered.

Declarations

Ethics approval The study was conducted following the ethical principles regarding human experimentation (Helsinki Declaration) and approved by the Ethics Committee of the Moscow State University of Psychology and Education.

Consent to participate All participants signed the informed consent before the experiment.

Conflict of interests We have no conflicts of interest to disclose.

References

Anderson, B. A. (2016). The attention habit: How reward learning shapes attentional selection. *Year in Cognitive Neuroscience, 1369*, 24–39. https://doi.org/10.1111/nyas.12957

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*, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Attard-Johnson, J., Ó Ciardha, C., & Bindemann, M. (2019). Comparing methods for the analysis of pupillary response. *Behavior Research Methods, 51*(1), 83-95. https://doi.org/10.3758/s13428-018-1108-6

Averbeck, B. B. (2015). Theory of Choice in Bandit, Information Sampling and Foraging Tasks. *PLoS Computational Biology, 11*(3), e1004164. https://doi.org/10.1371/journal.pcbi.1004164

Barthelme, S. (2019). eyelinker: Import ASC Files from EyeLink Eye Trackers. from https://cran.r-project.org/web/packages/eyelinker/index.html

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding Advantageously Before Knowing the Advantageous Strategy. *Science, 275*(5304), 1293. https://doi.org/10.1126/science.275.5304.1293

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics, 29*(4), 1165–1188.

Berlyne, D. E. (1966). Curiosity and exploration. *Science, 153*(3731), 25–33.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences, 8*(12), 539–546. https://doi.org/10.1016/j.tics.2004.10.003

Brevers, D., Noel, X., Bechara, A., Vanavermaete, N., Verbanck, P., & Kornreich, C. (2015). Effect of Casino-Related Sound, Red Light and Pairs on Decision-Making During the Iowa Gambling Task. *Journal of Gambling Studies, 31*(2), 409–421. https://doi.org/10.1007/s10899-013-9441-2

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience, 14*(11), 1462-U1140. https://doi.org/10.1038/nn.2925

Cavanagh, J. F., Wiecki, T. V., Kochar, A., & Frank, M. J. (2014). Eye Tracking and Pupillometry Are Indicators of Dissociable Latent Decision Processes. *Journal of Experimental Psychology-General, 143*(4), 1476–1488. https://doi.org/10.1037/a0035813

Cogliati Dezza, I., Yu, A. J., Cleeremans, A., & Alexander, W. (2017). Learning the value of information and reward over time when solving exploration-exploitation problems. *Scientific Reports, 7*(1), 16919. https://doi.org/10.1038/s41598-017-17237-w

Cohen, M. X., & van Gaal, S. (2013). Dynamic interactions between large-scale brain networks predict behavioral adaptation after perceptual errors. *Cerebral Cortex, 23*(5), 1061–1072. https://doi.org/10.1093/cercor/bhs069

Critchley, H. D., Tang, J., Glaser, D., Butterworth, B., & Dolan, R. J. (2005). Anterior cingulate activity during error and autonomic response. *NeuroImage, 27*(4), 885–895. https://doi.org/10.1016/j.neuroimage.2005.05.047

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

de Gee, J. W., Knapen, T., & Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences, 111*(5), E618. https://doi.org/10.1073/pnas.1317557111

Dudschig, C., & Jentzsch, I. (2009). Speeding before and slowing after errors: Is it all just strategy? *Brain Research, 1296*, 56–62. https://doi.org/10.1016/j.brainres.2009.08.009

Dyson, B. J., & Quinlan, P. T. (2003). Feature and conjunction processing in the auditory modality. *Perception & Psychophysics, 65*(2), 254–272. https://doi.org/10.3758/BF03194798

Egner, T. (2007). Congruency sequence effects and cognitive control. *Cognitive, Affective, & Behavioral Neuroscience, 7*(4), 380–390. https://doi.org/10.3758/CABN.7.4.380

Ellerby, Z. W., & Tunney, R. J. (2017). The Effects of Heuristics and Apophenia on Probabilistic Choice. *Advances in Cognitive Psychology, 13*(4), 280–295. https://doi.org/10.5709/acp-0228-9

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science, 306*(5703), 1940–1943. https://doi.org/10.1126/science.1102941

Gaffan, E. A., & Davies, J. (1981). The role of exploration in win-shift and win-stay performance on a radial maze. *Learning and Motivation, 12*(3), 282–299. https://doi.org/10.1016/0023-9690(81)90010-2

Garren, S. (2019). jmuOutlier: permutation tests for nonparametric statistics. R package version 2.2. from https://CRAN.R-project.org/package=jmuOutlier

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive Affective & Behavioral Neuroscience, 10*(2), 252–269. https://doi.org/10.3758/Cabn.10.2.252

Goudriaan, A. E., Oosterlaan, J., de Beurs, E., & van den Brink, W. (2005). Decision making in pathological gambling: A comparison between pathological gamblers, alcohol dependents, persons with Tourette syndrome, and normal controls. *Cognitive Brain Research, 23*(1), 137–151. https://doi.org/10.1016/j.cogbrainres.2005.01.017

Guttel, E., & Harel, A. (2005). Matching Probabilities: The Behavioral Law and Economics of Repeated Behavior. *U. Chi. l. Rev., 72*, 1197.

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest. *Journal of Statistical Software, 59*(9), 32. https://doi.org/10.18637/jss.v059.i09

Hershman, R., & Henik, A. (2019). Dissociation Between Reaction Time and Pupil Dilation in the Stroop Task. *Journal of Experimental Psychology-Learning Memory and Cognition, 45*(10), 1899–1909. https://doi.org/10.1037/xlm0000690

Higgins, J. J. (2004). *An introduction to modern nonparametric statistics*: Brooks/Cole Pacific Grove, CA.

Ivan, V. E., Banks, P. J., Goodfellow, K., & Gruber, A. J. (2018). Lose-Shift Responding in Humans Is Promoted by Increased Cognitive Load. *Frontiers in Integrative Neuroscience, 12*(9). https://doi.org/10.3389/fnint.2018.00009

Jepma, M., Beek, E. T. 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 psychopharmacological study. *Frontiers in human neuroscience, 4*https://doi.org/10.3389/Fnhum.2010.00170

Jepma, M., & Nieuwenhuis, S. (2011). Pupil Diameter Predicts Changes in the Exploration-Exploitation Trade-off: Evidence for the Adaptive Gain Theory. *Journal of Cognitive Neuroscience, 23*(7), 1587–1596. https://doi.org/10.1162/jocn.2010.21548

Joshi, S., & Gold, J. I. (2020). Pupil Size as a Window on Neural Substrates of Cognition. *Trends in Cognitive Sciences, 24*(6), 466–480. https://doi.org/10.1016/j.tics.2020.03.005

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron, 89*(1), 221–234. https://doi.org/10.1016/j.neuron.2015.11.028

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*(9), 697–720. https://doi.org/10.1037/0003-066X.58.9.697

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental Effects and Individual Differences in Linear Mixed Models: Estimating the Relationship between Spatial, Object, and Attraction Effects in Visual Attention. *Frontiers in psychology, 1*(238). https://doi.org/10.3389/fpsyg.2010.00238

Koenig, S., Uengoer, M., & Lachnit, H. (2018). Pupil dilation indicates the coding of past prediction errors: Evidence for attentional learning theory. *Psychophysiology, 55*(4), ARTN e13020. https://doi.org/10.1111/psyp.13020

Kolling, N., Behrens, T. E. J., Wittmann, M. K., & Rushworth, M. F. S. (2016). Multiple signals in anterior cingulate cortex. *Current Opinion in Neurobiology, 37*, 36–43. https://doi.org/10.1016/j.conb.2015.12.007

Kozunova, G., Voronin, N., Venidiktov, V., & Stroganova, T. (2018). Reinforcement Learning: a Role of Immediate Feedback and Internal Model. *Zhurnal Vysshei Nervnoi Deyatelnosti Imeni I.P. Pavlova, 68*(5), 602–613. https://doi.org/10.1134/S0044467718050076

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 1*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Laeng, B., Orbo, M., Holmlund, T., & Miozzo, M. (2011). Pupillary Stroop effects. *Cognitive Processing, 12*(1), 13–21. https://doi.org/10.1007/s10339-010-0370-z

Lavin, C., San Martin, R., & Jubal, E. R. (2014). Pupil dilation signals uncertainty and surprise in a learning gambling task. *Frontiers in behavioral neuroscience, 7*, Artn 218. https://doi.org/10.3389/Fnbeh.2013.00218

Lenth, R. V. (2021). emmeans: estimated marginal means, aka least-squares means. R package version 1.6. 0. from https://CRAN.R-project.org/package=emmeans

Lin, H., Saunders, B., Hutcherson, C. A., & Inzlicht, M. (2018). Midfrontal theta and pupil dilation parametrically track subjective conflict (but also surprise) during intertemporal choice. *NeuroImage, 172*, 838–852. https://doi.org/10.1016/j.neuroimage.2017.10.055

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods, 50*(1), 94–106. https://doi.org/10.3758/s13428-017-1007-2

Murphy, P. R., van Moort, M. L., & Nieuwenhuis, S. (2016). The Pupillary Orienting Response Predicts Adaptive Behavioral Adjustment after Errors. *PLoS ONE, 11*(3), e0151763. https://doi.org/10.1371/journal.pone.0151763

O'Connell, R. G., Dockree, P. M., Robertson, I. H., Bellgrove, M. A., Foxe, J. J., & Kelly, S. P. (2009). Uncovering the Neural Signature of Lapsing Attention: Electrophysiological Signals Predict Errors up to 20 s before They Occur. *Journal of Neuroscience, 29*(26), 8604–8611. https://doi.org/10.1523/jneurosci.5967-08.2009

Payzan-LeNestour, E., & Bossaerts, P. (2012). Do not bet on the unknown versus try to find out more: estimation uncertainty and "unexpected uncertainty" both modulate exploration. *Frontiers in Neuroscience, 6*. https://doi.org/10.3389/fnins.2012.00150

Poe, G. R., Foote, S., Eschenko, O., Johansen, J. P., Bouret, S., Aston-Jones, G., . . . Sara, S. J. (2020). Locus coeruleus: a new look at the blue spot. *Nature Reviews Neuroscience, 21*(11), 644-659. https://doi.org/10.1038/s41583-020-0360-9

Preciado, D., Munneke, J., & Theeuwes, J. (2017). Mixed signals: The effect of conflicting reward- and goal-driven biases on selective attention. *Attention, Perception, & Psychophysics, 79*(5), 1297–1310. https://doi.org/10.3758/s13414-017-1322-9

Preuschoff, K., Hart, B. M., & Einhauser, W. (2011). Pupil dilation signals surprise: evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience, 5*, Unsp 115. https://doi.org/10.3389/Fnins.2011.00115

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Richer, F., & Beatty, J. (1987). Contrasting Effects of Response Uncertainty on the Task-Evoked Pupillary Response and Reaction-Time. *Psychophysiology, 24*(3), 258–261. https://doi.org/10.1111/j.1469-8986.1987.tb00291.x

Satterthwaite, T. D., Green, L., Myerson, J., Parker, J., Ramaratnam, M., & Buckner, R. L. (2007). Dissociable but inter-related systems of cognitive control and reward during decision making: Evidence from pupillometry and event-related fMRI. *NeuroImage, 37*(3), 1017–1031. https://doi.org/10.1016/j.neuroimage.2007.04.066

Saunders, B., Lin, H., Milyavskaya, M., & Inzlicht, M. (2017). The emotive nature of conflict monitoring in the medial prefrontal cortex. *International Journal of Psychophysiology, 119*, 31–40. https://doi.org/10.1016/j.ijpsycho.2017.01.004

Sayfulina, K., Kozunova, G., Medvedev, V., Rytikova, A., & Chernyshev, B. (2020). Decision making under uncertainty: exploration and exploitation. *Journal of Modern Foreign Psychology, 9*(2), 93–106. https://doi.org/10.17759/jmfp.2020090208

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology, 55*, 7–14. https://doi.org/10.1016/j.conb.2048.11.003

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H. B., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *eLife, 8*, e41703. https://doi.org/10.7554/eLife.41703

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making, 15*(3), 233–250. https://doi.org/10.1002/bdm.413

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function. *Neuron, 79*(2), 217–240. https://doi.org/10.1016/j.neuron.2013.07.007

Stuart, A., Ord, J. K., & Arnold, S. (1999). *Kendall's advanced theory of statistics. vol. 2a: Classical inference and the linear model*. London: Arnold.

Sutton, R. S., & Barto, A. G. (1999). Reinforcement Learning. *Journal of Cognitive Neuroscience, 11*(1), 126–134.

Tibon, R., & Levy, D. A. (2015). Striking a balance: analyzing unbalanced event-related potential data. *Frontiers in psychology, 6*(555). https://doi.org/10.3389/fpsyg.2015.00555

Tukey, J. (1977). Exploratory data analysis (Vol. 2, pp. 131–160). *Reading, PA: Addison-Wesley*.

Unsworth, N., & Robison, M. K. (2016). Pupillary correlates of lapses of sustained attention. *Cognitive, Affective, & Behavioral Neuroscience, 16*(4), 601–615. https://doi.org/10.3758/s13415-016-0417-4

Unturbe, J., & Corominas, J. (2007). Probability matching involves rule-generating ability: A neuropsychological mechanism dealing with probabilities. *Neuropsychology, 21*(5), 621–630. https://doi.org/10.1037/0894-4105.21.5.621

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications, 8*(1), 14637. https://doi.org/10.1038/ncomms14637

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., & Aston-Jones, G. (1999). The role of locus coeruleus in the regulation of cognitive performance. *Science, 283*(5401), 549–554. https://doi.org/10.1126/science.283.5401.549

Van Slooten, J. C., Jahfari, S., Knapen, T., & Theeuwes, J. (2018). How pupil responses track value-based decision-making during and after reinforcement learning. *PLoS computational biology, 14*(11). https://doi.org/10.1371/journal.pcbi.1006632

Vossen, H., Van Breukelen, G., Hermens, H., Van Os, J., & Lousberg, R. (2011). More potential in statistical analyses of event-related potentials: A mixed regression approach. *International Journal of Methods in Psychiatric Research, 20*(3), e56–e68. https://doi.org/10.1002/mpr.348

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14*(1), 101–118. https://doi.org/10.1111/1467-6419.00106

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

Wessel, J. R., Danielmeier, C., & Ullsperger, M. (2011). Error Awareness Revisited: Accumulation of Multimodal Evidence from Central and Autonomic Nervous Systems. *Journal of Cognitive Neuroscience, 23*(10), 3021–3036. https://doi.org/10.1162/jocn.2011.21635

Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Current Opinion in Behavioral Sciences, 38*, 49–56. https://doi.org/10.1016/j.cobeha.2020.10.001

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

Zajkowski, W. K., Kossut, M., & Wilson, R. C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. *eLife, 6*, e27430. https://doi.org/10.7554/eLife.27430

Zenon, A. (2019). Eye pupil signals information gain. *Proceedings of the Royal Society B-Biological Sciences, 286*(1911), Artn 20191593. https://doi.org/10.1098/Rspb.2019.1593