**Title:** Pupil correlates of decision variables in mice playing a competitive mixed-strategy game

**Authors:** Hongli Wang1, Heather K. Ortega1, Huriye Atilgan2, Cayla E. Murphy2, Alex C. Kwan2,3

1Interdepartmental Neuroscience Program,

2Department of Psychiatry,

3Department of Neuroscience,

Yale University School of Medicine, New Haven, Connecticut, 06511, USA

Correspondence: alex.kwan@yale.edu

**Abstract**

Animals are flexible decision makers and can adapt to competitive pressure. More specifically, in a competitive task involving two players where the outcome is contingent on choices of both players, the animal must continually adjust or risk being exploited and losing out on rewards. In this study, we demonstrate that head-fixed mice can be trained to play the iterative competitive game ‘matching pennies’ against a virtual computer opponent. Comparing between a variety of learning algorithms, we show that the animals’ performance is well described by a hybrid computational model that includes Q-learning and choice kernel. To understand the involvement of neuromodulatory mechanisms, we measure fluctuations in pupil size and use multiple linear regression to relate the trial-by-trial transient pupil responses to decision-related variables. The analysis revealed that pupil responses are modulated by observable variables, including choice and outcome, as well as latent variables for value updating, but not action selection. Finally, we contrast these results from matching pennies against behavioral and pupil data obtained from a two-armed bandit task. Collectively, these results establish a paradigm for studying competitive decision-making in head-fixed mice and, through pupillary measurements, provide insights into the potential role of arousal-linked neuromodulation in the decision process.

**INTRODUCTION**

Animals learn from the outcomes of their past actions. The decision-making process can be casted in the framework of reinforcement learning (Sutton and Barto, 1998), which provides a quantitative approach to characterize how animals choose among multiple options based on the prior experience. This approach, when applied to rodents and combined with powerful molecular, genetic, electrophysiological, and imaging methods, has yielded novel insights into the neural circuits involved in reward-based learning (Ito and Doya, 2009; Sul et al., 2010; Sul et al., 2011; Tai et al., 2012; Bari et al., 2019; Groman et al., 2019; Hattori et al., 2019). However, to date most studies in rodents relied on tasks involving a pure strategy, where there is always a particular action that is optimal in a given situation. Pure strategy is possible for some but not all tasks. For instance, in a two-player game (e.g., rock-paper-scissors) where the outcome depends on both the animal and an opponent, pure strategies are inadequate because an opponent can predict tendencies and exploit to win. Instead, under competitive pressure, the animal should adopt a mixed strategy, in which two or more pure strategies are chosen probabilistically. Analyses of behavior under such conditions fall into the purview of game theory (Camerer, 2003), and provide a unique window into the social and adaptive aspects of decision-making (Lee, 2008).

Research in humans and non-human primates has identified numerous brain regions contributing to decision-making during two-player games. For example, a functional imaging study in humans showed widespread representation of reward signals in the brain during multiple types of games against computerized opponents (Vickery et al., 2011). Signals related to choices and mentalizing the opponent’s actions were localized to several brain regions including the prefrontal cortex (Hampton et al., 2008; Vickery et al., 2011). Electrophysiological recordings in macaque monkeys demonstrated spiking activity patterns that suggest neural representations of various decision variables in prefrontal cortical regions, lateral intraparietal cortex, and amygdala (Barraclough et al., 2004; Seo et al., 2007, 2009; Chang et al., 2013; Haroush and Williams, 2015; Dal Monte et al., 2020). However, there have only been a few reports of rodents engaging in two-player games (Tervo et al., 2014; Wood et al., 2016), and the associated neural correlates are less clear.

Beyond neural activity, the neuromodulator norepinephrine (NE) may be particularly relevant for animals to achieve optimal performance during two-player games. A large body of literature has linked central noradrenergic tone to behavioral flexibility (Aston-Jones and Cohen, 2005; Bouret and Sara, 2005). In one pioneering study, Tervo et al. taught rats to play a matching pennies game with a computer opponent (Tervo et al., 2014). By manipulating neural activity in locus coeruleus (LC), they showed that elevating central NE tone can suppress firing in the anterior cingulate cortex, which in turn reduces the influence of reinforcement history and promotes stochastic behavior. A different way to study neuromodulatory tone is to measure pupil size, which is often treated as a readout of NE levels in the neocortex (Gilzenrat et al., 2010; Joshi et al., 2016; Reimer et al., 2016). Studies of pupillary dynamics during tasks further support the idea that NE is important for flexible decisions. For instance, the baseline pupil size is shown to correlate with biases in the explore-exploit trade-off and attentional set shifting (Jepma and Nieuwenhuis, 2011; Pajkossy et al., 2017). The trial-by-trial transient change in pupil size is reported to associate with many task-relevant variables including upcoming choice, expected outcome, values of the choices, and uncertainties (Hess and Polt, 1960; Qiyuan et al., 1985; de Gee et al., 2014; Van Slooten et al., 2018). Given these prior results, it seems likely that pupil fluctuations can be leveraged to study the neuromodulatory mechanisms underlying adaptive action selection during competitive two-player games.

We have two goals for the current study. First, we want to know if head-fixed mice can engage and compete proficiently in a two-player game. Second, we want to characterize pupil fluctuations to gain insights into the role of neuromodulation in decision-making under competitive pressure. To this end, we design a behavioral paradigm for a head-fixed mouse to play iterative matching pennies against a computer-controlled virtual opponent. We find that mice can perform at a level close to the optimal reward rate by exhibiting choice behavior consistent with a mix of reinforcement learning and choice perseveration. We show transient, within-trial changes in pupil size that are associated with the choice, outcome, and latent variables for value updating. Finally, we contrast these choice behavior and pupil fluctuations recorded during matching pennies with those quantified in a more widely used decision-making paradigm for mice, the two-armed bandit task.

**MATERIALS AND METHODS**

**Animals**

All animal procedures were conducted in accordance with procedures approved by the Institutional Animal Care and Use Committee at Yale University. Adult male C57BL/6J mice (>P56; #000664, Jackson Laboratory) were used for all experiments. Mice were housed in groups of 3-5 animals with 12-hr-dark/12-hr-light cycle control (lights off at 19:00).

**Surgical Procedures**

Anesthesia was induced with 2% isoflurane in oxygen before the surgery. The isoflurane was lowered to 1–1.2% during the surgical procedures. The mouse was placed on a water-circulating heating pad (TP-700, Gaymar Stryker) in a stereotaxic frame (David Kopf Instruments). After injecting carprofen (5 mg/kg, s.c.; #024751, Butler Animal Health) and dexamethasone (3 mg/kg, s.c.; Dexaject SP, #002459, Henry Shein Animal Health), the scalp was removed to expose the skull. A custom-made stainless-steel head plate (eMachineShop) was glued onto the skull with MetaBond (C&B, Parkell, Inc.). Carprofen (5 mg/kg, s.c.) was injected each day for the following three days. Mice were given 7 days to recover from the surgery before any behavioral training.

**Behavioral setup**

The training apparatus was based on a previous design from our prior studies (Siniscalchi et al., 2016; Siniscalchi et al., 2019). Detailed instruction to construct the apparatus is available at <https://github.com/Kwan-Lab/behavioral-rigs>. The mouse with a head plate implant was fixed to a stainless-steel holder (eMachineShop). The animal sat inside an acrylic tube (8486K433; McMaster-Carr), which limited gross movements though allowed postural adjustments. A lick port with two lick spouts was positioned in front of the subject. The spouts were constructed with 20-gauge stainless-steel needles. Contact with the lick spout – which is how the animal indicates its choices – was detected via wires that were soldered onto the spout that connect to a battery-powered lick detection electronic circuit. Output signals from the circuit were sent to a computer via a data acquisition unit (USB-201, Measurement Computing) and logged by the Presentation software (Neurobehavioral Systems, Inc.). Water delivery via the lick spouts was controlled independently for each spout by a solenoid fluid valve (MB202-V-A-3-0-L-204; Gems Sensors & Controls). The amount of water was calibrated to ~4 μL per pulse by adjusting the duration of the electrical pulse sent by the Presentation software via a second data acquisition unit (USB-201, Measurement Computing). Two speakers (S120, Logitech) were placed in front of the mouse to play the sound cue. The whole setup was placed inside an audio-visual cart with walls lined with soundproof acoustic foams (5692T49, McMaster-Carr). A monochrome camera (GigE G3-GM11-M1920, Dalsa) with a 55 mm telecentric lens (TEC-55, Computar) was aimed at the right eye. Video was acquired at 20 Hz. A dimmable, white light source (LT-T6, Aukey) was used to provide ambient light, such that the baseline pupil size was moderate and fluctuations around the baseline is detectable. The computer running the Presentation software would send TTL pulses to a USB data acquisition device (USB-201; Measurement Computing) connected to the computer controlling the camera. This computer for the camera would run a custom script written in MATLAB 9.7 (MathWorks) that logged the timing of the TTL pulses such that the behavioral log files generated by the Presentation software may be aligned to the video recordings. In a small subset of experiments, we captured videos from both left and right eyes by mounting two identical camera systems on both sides of the animal.

**Behavioral training – Shaping**

All of the procedures for initial shaping as well as the final matching pennies and two-armed bandit tasks were written using the scripting language in the Presentation software. The animals were fluid restricted. Water was provided during the one behavioral session daily. On the days when the animal was not trained (typically 1 day a week), a water bottle was placed in the home cage for 5 min of ad libitum consumption. All animals underwent the same two shaping phases before training on matching pennies or two-armed bandit task. For phase 1 (2 days), the animal was habituated to the apparatus. They may lick either spout. A water reward would be delivered for every lick in the corresponding spout, as long as a minimum of 1 s has occurred since the last reward. The session would terminate after the animal acquired 100 rewards. For phase 2 (~4 weeks), the animal was introduced to the trial structure and learned to suppress impulsive licks. At the start of each trial, a 5 kHz sound cue lasting for 0.2 s was played. From the onset of the sound cue, the mouse has a window of 2 s to make a response by licking either of the spouts. If a lick was detected during the response window, a water reward would be delivered in the corresponding spout and there is a fixed 3 s period for consumption. From the end of the consumption window or the end of response window if no lick was detected, an inter-trial interval (ITI) began. The duration of the ITI in seconds was drawn from a truncated exponential distribution with λ = 1/3 and boundaries of 1 and 5. If the animal emit one or more lick during the ITI, then additional time drawn again from the same truncated exponential distribution would be appended to the ITI. If the mouse licked again during the appended time, yet another additional time would be appended, up to a total of 5 draws including the initial draw. When the ITI ends, a new trial would begin. This trial timing is the same as what will be used for matching pennies and two-armed bandit task. The animal would advance to the next stage of training when the average number of draws per ITI was lower than 1.2 for three consecutive sessions. Note that the goal for the shaping was to habituate and introduce lick suppression. Although the mouse could theoretically get water from either spout, the animal tended to favor heavily one spout during the shaping procedures.

**Behavioral training – Matching pennies**

For matching pennies, the mouse played against a virtual opponent in the form of a computer agent. At the start of each trial, the agent made a choice (left or right). If the mouse selected the same choice as the computer, a water reward would be delivered in the corresponding spout. Otherwise, no reward was presented. Importantly, the computer agent was designed to provide competitive pressure by acting according to prediction of the animal’s choices. Specifically, it was programmed to be the same as ‘algorithm 2’ (Barraclough et al., 2004; Lee et al., 2004) or ‘competitor 1’ (Tervo et al., 2014) in previous studies. Briefly, the agent had a record of the mouse’s entire choice and reward history within the current session. The agent calculated the conditional probabilities that the animal would choose left given sequences of the preceding N choices (N = 0 – 4) and sequences of preceding N choice-outcome combinations (N = 1 – 4). The binomial test was used to test each of the 9 conditional probabilities against the null hypothesis that the mouse would choose left with a probability of 0.5. If none of the null hypothesis was rejected, the agent would randomly choose either target with equal probabilities. If one or more hypotheses were rejected, the agent would generate the counter choice with the statistically significant conditional probability that was farther away from 0.5. A session would terminate automatically when no response was logged for 10 consecutive trials. When an animal reached a 40% reward rate for 3 consecutive sessions (~4 weeks), then its performance was considered stable and the subsequent sessions were included in the following analysis.

**Behavioral training – Two-armed bandit**

For the two-armed bandit task, the two options (left and right) were associated with different reward probabilities. In each trial, when an animal chose an option, a reward was delivered stochastically based on the assigned reward probability. In our implementation, there were two sets of reward probabilities: 0.7:0.1 and 0.1:0.7. A set of reward probabilities were maintained across a block of trials. Within a block, once the mouse had chosen the option with high reward probability (hit trials) for ten times, then on any given trial there was a probability of 1/11 that the reward probabilities would change. Thus, the number of trials in a block after ten hits followed a geometric distribution with = 11. There were no explicit cues to indicate a block switch, therefore the animal had to infer the current situation through experience.

**Preprocessing of behavioral data**

A total of 107 sessions from 9 mice were included in the study, including 81 sessions from 5 mice trained on matching pennies and 26 sessions from 4 mice trained on the two-armed bandit task. All of the sessions contained both behavioral and pupillometry data. For behavior, the log file saved by the Presentation software contained timestamps for all events that occurred during a session. Analyses of the behavioral data were done in MATLAB. For matching pennies, towards the end of each session, the animals tended to select the same option for around 30 trials before ceasing to respond. To avoid these repetitive trials in the analyses, for each session, the running 3-choice entropy (see below) of a 30-trial window was calculated, and the MATLAB function *ischange* was used to fit with a piecewise linear function. The trial when the fitted function fell below a value of 1 was identified as the ‘last trial’, and all subsequent trials were discarded for the analysis. In some cases, the performance recovered after the detected last trial to a value greater than 1, then the entire session was used for analysis. In some cases, the fitted function did not fall below a value of 1 and no ‘last trial’ was detected, then the entire session was used for analysis.

**Analysis of behavioral data – Entropy**

To quantify the randomness in the animals’ choices, the 3-choice entropy of the sequence of the choices was calculated by:

where is the frequency of occurrence of a 3-choice sequence in a session. Because there were 2 options to choose from, there were 23 = 8 potential sequences possible. The maximum value for entropy is 3 bits.

**Analysis of behavioral data – Computational models**

To quantify the choice behavior, we considered 5 computational models. The primary model used in the paper is a Q-learning with forgetting model plus a choice kernel (FQ\_RPE\_CK) (Katahira, 2015; Wilson and Collins, 2019). On trial , for a choice that leads to an outcome , the action value associated with an action is updated by:

where is the learning rate, is the forgetting rate for the unchosen action. In our task, there are two options, so . For the outcome, = 1 for reward, 0 for no reward. Moreover, to capture the animal’s tendency to make decisions based purely on previous choices, there are choice kernels that were updated by:

where is the choice-kernel learning rate. For action selection, the probability to choose action on trial is given by a softmax function:

where and are the inverse temperature parameters for action values and choice kernels respectively.

We compared the FQ\_RPE\_CK model against 4 other models. For the win-stay-lose-switch model (WSLS), the probability to choose action on trial is given by:

where is the probability that the animal followed the win-stay-lose-switch strategy.

For the Q-learning model (Q\_RPE), the action value is updated by:

and the probability to choose action at trial was then given by:

For the forgetting Q-learning model (FQ\_RPE), the action values were updated by equation (1), and the probability to choose action on trial is given by equation (7).

For the differential Q-learning model (DQ\_RPE) (Caze and van der Meer, 2013; Katahira, 2018), the action value is updated by:

where and are the learning rates for rewarded and unrewarded trials, respectively. The probability to choose action on trial was given by equation (7).

**Analysis of behavioral data – Model fitting and comparison**

To fit the computational models to the behavioral data, for each subject, the sequence of choices and rewards were concatenated across sessions. Each model was fitted to the data using a maximum-likelihood algorithm implemented with the *fmincon* function in MATLAB, with the constraints , , and . These fits also yielded latent decision variables such as action values and choice kernels that would be used for the subsequent multiple linear regression analyses. For model comparison, we calculated the Bayesian information criterion (BIC) for each of the model fits.

**Preprocessing of pupillometry data**

To extract the coordinates of the pupil from the video recordings, we used DeepLabCut 2.0 (Mathis et al., 2018; Nath et al., 2019), ran on Jupyter Notebook on Google’s cloud servers. A small subset of the video frames was manually analyzed, with the experimenter annotating 5 labels including the central, uppermost, leftmost, lowermost, and rightmost points of the pupil. The annotated frames were fed to DeepLabCut to teach a deep neural network to analyze the remainder of the video to produce the 5 labels. From the labels, the pupil diameter was computed by taking the distance between the leftmost and rightmost labels. We did not use the other labels, because the estimates of the lowermost points were unstable, sometimes jumping in consecutive frames due to interference from the lower eyelid. The pupil diameter signal was further processed through a 4 Hz lowpass filter with the MATLAB function *lowpass*. Then a threshold was used to delete any frames with unreasonably large or small values using the MATLAB function *isoutlier*. Using a 10-minute moving window to account for drift over a session, the signal was converted to z-score. Finally, we calculated the pupil response for each trial by subtracting the instantaneous z-score from -3 to 5 s relative to cue onset by the baseline z-score, which was defined as the mean z-score from -1 to 0 s relative to cue onset.

**Analysis of pupil data – Multiple linear regression**

To determine how pupil responses may relate to choices and outcomes, we used multiple linear regression:

where is the pupil response at time in trial , are the choices made on the next trial, the current trial, the previous trial, and the trial before the previous trial, respectively, are the outcomes for the next trial, the current trial, the previous trial, and the trial before the previous trial, respectively, ,…, are the regression coefficients, and is the error term. Choices were dummy-coded as 0 for left responses and 1 for right responses. Outcomes were dummy-coded as 0 for no-reward and 1 for reward. For the last 2 predictors in Eq. 9, is the average reward over the previous 20 trials, given by the equation:

The term indicates the normalized cumulative reward during the session, calculated by:

where denotes the current trial number and is the total number of trials in the session.

To determine how pupil responses may relate to latent decision variables for action selection, we used multiple linear regression:

where  and denote the action values of the left and right choices in trial , respectively, is the value of the action chosen in trial , and are the choice kernels of the left and right choices in trial , respectively, is the choice kernels of the action chosen in trial .

To determine how pupil responses may relate to latent decision variables for value updating, we used multiple linear regression:

where is the reward prediction error (RPE), and is the error used to update the choice kernels, or choice kernel error (CKE).

For each session, the regression coefficients were determined by fitting the equations to data using the MATLAB function *fitlm*. The fit was done in 100-ms time bins that span from -3 to 5 s relative to cue onset, using mean pupil response within the time bins. To summarize the results, for each predictor, we calculated the proportion of sessions in which the regression coefficient was different from zero (*P* < 0.01). To determine if the proportion was significantly different from chance, we performed a chi-square test against the null hypothesis that there is a 1% probability that a given predictor is mischaracterized as significant by chance in a single session.

**Code accessibility**

The data and code that support the findings of this study will be made publicly available at <http://github.com/Kwan-Lab>.

**RESULTS**

**Mice played matching pennies against a computer opponent**

We trained head-fixed mice to play matching pennies against a computer opponent (**Fig. 1A**). In this iterative version of matching pennies, each trial the mouse and the computer would choose left or right. If the actions match, the mouse receives a water reward; otherwise, the mouse receives nothing. The corresponding payoff matrix is shown in **Figure 1B**. This game was challenging for the mouse because the computer opponent had access to the complete history of choices and rewards over the session and was programmed to predict the mouse’s next choice in order to make the counter action (see Materials and Methods; same as ‘algorithm 2’ in (Lee et al., 2004) and ‘competitor 1’ in (Tervo et al., 2014)). **Figure 1C** shows the trial structure. A 0.2 s, 5 kHz sound cue indicated the start of the trial. Within a 2-s response window, the mouse could make a directional tongue lick to the left or right spout to indicate its choice. Based on the animal’s and computer’s choices and the payoff matrix, the animal might receive a water reward after the response. To minimize pre-cue licks, the intertrial interval was drawn from a truncated exponential distribution and could be extended if the mouse could not suppress licking (see Materials and Methods). Mice were trained daily and, on average, took about 4 weeks to follow the trial timing to suppress pre-cue licking, and then another 4 weeks playing matching pennies to achieve a performance threshold of 40% reward rate for 3 consecutive sessions. All the data and analyses presented in the paper came from sessions after the threshold was attained.

The data set for matching pennies included 81 behavioral sessions with concurrent pupil measurements (n = 5 mice). For matching pennies, the Nash equilibrium based on game theory indicates that optimally the animal should choose left and right with equal probabilities, which would yield a long-run reward rate of 50%. In an example session, plotting the choices and rewards for a mouse and the choices for the computer showed that the animal exhibited a great degree of stochasticity in their choice pattern (**Fig. 1D**). This could be seen more clearly by looking at the cumulative occurrences of various plausible 3-choice sequences (**Fig. 1E**). Although there was sometimes a slight preference for certain patterns, particularly early in the session, throughout most of the sessions the animal employed continually the different choice sequences. Across all sessions, animals performed 501±13 trials per session (mean±s.e.m.; **Fig. 1F**). The entropy for 3-choice sequences, a measure of the stochasticity in choices, was 2.87±0.01 bits, close to the theoretical upper bound of 3 bits. Consequently, the computer was only mildly effective at predicting the mouse’s choices, and the mice earned an average reward rate of 44.0±0.4%. We note that although the reward rate compared favorably with the optimal reward rate of 50%, the difference was statistically significant (*P* = 2 x 10-20, one-sample one-tailed t-test). Finally, mice had stereotypical response times that were similar for left and right choices (**Fig. 1G**). Altogether, these results demonstrate that head-fixed mice can play matching pennies against a computer opponent. Given that humans and macaques likewise play matching pennies imperfectly and not at Nash equilibrium (Erev and Roth, 1998a; Lee et al., 2004), here we found that reward rate was decent but suboptimal for mice, suggesting that the animals had certain residual tendencies that were exploited by the computer opponent.

|  |
| --- |
|  |
| **Figure 1. Mouse’s behavior in matching pennies game.**  (A) A schematic illustration of the competitive game. Head-fixed mouse makes a left or right choice by licking the corresponding port. A computer monitors the mouse’s choices and outcomes, generating a left or right choice every trial. (B) Payoff matrix of the game. The mouse will receive a water reward only if it chooses the same choice as the computer does in the same trial. (C) Trial timings: the mouse must wait for a go cue before licking the ports. The outcome is presented immediately after the response. A random intertrial interval follows the outcome. (D) An example session. Top: the mouse’s choice and outcome. Black bar indicates a reward in the trial. The average reward rate is 52.1%. Blue vertical lines on top indicate right choices. Red vertical lines below indicate left choices. Bottom: the computer’s choices. Color codes are the same as the top panel. (E) Cumulative number of 3-choice patterns in the same session as (E). Different lines represent different 3-choice combination. (F) Summary of the behavior (81 sessions, 5 subjects). Left: the average trials performed every session is 501±13; Middle: the average entropy of the 3-choice sequences every session is 2.87± 0.01; Right: the average reward rate is 44.0% ± 0.4%.(G) The histogram of the response time. Left: the histogram of the response time in all trials; Middle: the histogram of the response time in trials which mice chose left; Right: the histogram of the response time in trials which mice chose right. |

**Animals’ behavior was captured by a hybrid model with reinforcement learning and choice kernel**

What is the strategy that characterizes the tendencies in the animal’s behavior? Previous studies in macaque monkeys found that reinforcement learning can account in part for the animals’ behavior in matching pennies (Lee et al., 2004; Seo et al., 2007). We therefore compared between a range of strategies: win-stay-lose-switch, and reinforcement learning algorithms including Q-learning (Q\_RPE), differential Q-learning (DQ\_RPE), and forgetting Q-learning (FQ\_RPE) (Ito and Doya, 2009) (see Materials and Methods). Fitting each model to the behavioral data, the Bayesian information criterion (BIC) values indicated that FQ\_RPE was most consistent with the observed choice behavior. To further enhance the FQ\_RPE algorithm, we note that mice exhibited serial choice dependence and sometimes favored picking the same choice in successive trial (i.e., **Figure 1E**). To capture perseverative behavior, we added choice kernels (Wilson and Collins, 2019) to create the FQ\_RPE\_CK algorithm. **Figure 2A** illustrates graphically the FQ\_RPE\_CK scheme (henceforth called the ‘hybrid model’), with reinforcement learning through FQ\_RPE and perseveration through choice kernel. Note that both the action values and choice kernels were updated on a trial-by-trial basis, and each had its own learning rate ( and ) and inverse temperature parameter ( and ; see Materials and Methods). A head-to-head comparison of all the models indicated that the hybrid FQ\_RPE\_CK model provided the most accurate fit to the behavioral data (**Fig. 2B**). For the remainder of the analyses, we will quantify the animal’s strategy using the hybrid model.

**Figure 2C** shows a representative session including the animal’s choices and outcomes as well as the latent variables, including action values (middle panel) and choice kernels (bottom panel), estimated by fitting the hybrid model. The model-estimated probability of choosing left (gray line, **Fig. 2C**) tracked the actual choice pattern (black line, **Fig. 2C**). To further validate the fit, we plotted the animal’s choice probability as a function of the model-estimated, weighted sum of action values and choice kernels, which was the crucial parameter used to simulate action selection each trial (**Fig. 2D**). This analysis showed that on most trials, animals selected with a weighted sum that was close to zero, i.e., with near equal probabilities of choosing left and right, making it difficult for the computer opponent to predict their choice. Moreover, the observed choice probability (circles, **Fig. 2D**) fitted well to the softmax function used for action selection (dashed line, **Fig. 2D**), except for deviations for very low or high weight sum values. Taken together, these analyses indicate that the mouse’s behavior during matching pennies can be quantified using a hybrid model including reinforcement learning and choice kernels.

|  |
| --- |
|  |
| **Figure 2. Reinforcement learning model with choice kernels reliably tracks animals’ behavior.**  (A) Schematic of the reinforcement learning model with choice kernels. denote the action value for left and right choices, respectively; denote the choice kernels for left and right choices, respectively. The animal makes a decision stochastically (softmax) based on the difference of the action values and choice kernels (Δ*Q*, Δ*K*), then compute the choice kernel error (CKE), which is the difference between the actual choice and the value of the choice kernel. The CKE is then used to update the choice kernels. While the outcome is used to calculate the RPE, which is used to update the action values. (B) Model comparison using BIC. The models were fit to different subjects, then the average of BICs over subjects were used to compare different models. (C) An example of the time course of the latent variables and predicted behavior. Top: Long vertical red and blue lines indicate rewarded left and right choices, respectively; short vertical red and blue lines indicate unrewarded left and right choices. Gray line shows the average probability to choose left, smoothed by a gaussian kernel. Black line shows the probability to choose left predicted by the model. Middle: the action values of left (red line) and right (blue line) choices estimated by the model; Bottom: the choice-autocorrelation factors of left (red) and right (blue) choices estimated by the model. (D) Psychometric curve fitted to the whole data set. Black histograms indicate the distribution of trials according to the weighted sum of the differences of action values (Δ*Q*) and the difference of choice kernels (Δ*K*). β, βK are the weights of action values and choice kernels, respectively. Dashed purple line shows the predicted probability to choose left according to the model; Purple dots show the probability to choose left computed from animals’ behavior. |

With the hybrid model, it is a logical question to ask how the two components – reinforcement learning and choice kernels – were balanced during matching pennies. Fits to behavioral data indicated that the choice kernel was weighed more strongly than action values during action selection (*β*K/(*β*K+*β*) = 0.75±0.17; **Fig. 3A**), although the ratio of learning rates suggests that the choice kernels were updated slower than the action values (*α*K/*α* = 0.22±0.10). The different dynamics of the two components was consistent with the example in **Figure 2C**, where the action values fluctuate more rapidly than the choice kernels. The overall exploration level could be inferred by the sum of the inverse temperatures (*β*K+*β* = 2.15± 0.17). To find out how varying the balance between reinforcement learning and choice kernel would affect the performance, we simulated performance in match pennies using a computer agent, varying *β*K/(*β*K+*β*) while fixing *α*K/*α* and *β*K+*β*. The simulations revealed that the reward rate was relatively stable if the computer agent used mostly reinforcement learning or a hybrid strategy (**Fig. 3B**). However, if the computer agent based its actions exclusively on choice kernels, the performance declined precipitously. It was intriguingly to see the *β*K/(*β*K+*β*) estimated from behavioral data lied around the threshold between these two conditions, indicating that the animals might have settled on a hybrid strategy that balanced a tradeoff between performance and effort.

|  |
| --- |
|  |
| **Figure 3. Parameters estimation and simulation results of a computational agent playing the matching pennies game****.**  (A) Parameters estimation of the actual animal behavior. Left panel: the relative weight of choice kernel. *β*K/(*β*+*β*K) = 0.75±0.09; Middle panel: the total exploration level. *β*+*β*K = 2.15±0.17; Right panel: relative learning rate of the choice kernel. *α*K/*α* = 0.22±0.10. (B) The performance of a computational agent playing the game. The agent adopted different values of *β*K/(*β*+*β*K), with the summation of inverse temperatures (*β*+*β*K) unchanged. The values of learning rate and summation of inverse temperature are set to be the median of the fitted value of the five animals. Black solid curve: the probability of stay as a function of different relative inverse temperature; Black dashed curve: average reward rate; Purple solid curve: entropy of three consecutive choices. Black solid, black empty, and purple dot: the probability of stay, average reward, and entropy of three consecutive choices of the computational agent whose all four parameters were set to be the median of the fitted parameters. |

**Pupillometry data were processed using DeepLabCut**

While the mouse played matching pennies, a camera was positioned to capture videos of the right eye (**Fig. 4A**). To extract pupil size from the video, we labeled a small subset of video frames manually, which were fed to the toolbox DeepLabCut (Mathis et al., 2018; Nath et al., 2019) to generate a deep convolutional network to extract automatically labels from the entire video. **Figure 4B** shows an example video frame with labels identified manually and by DeepLabCut. To quantify the performance of the automated procedure, we calculated the deviations between the manually selected and automatically estimated labels for the top, bottom, left, right, and center positions of the mouse’s pupil (**Fig. 4C - D**). The mean values of the deviations were close to zero, demonstrating that the estimates had little bias. To give an intuition into the pupil size fluctuation observed, for one session, we plotted the time course of the pupil diameter after z-score normalization (**Fig. 4E**). When aligned to select trial types, there were obvious task-related transients in the pupil size (lower panels, **Fig. 4E**). In this study, we were interested in the relation between pupil fluctuations and decision-related variables on a trial-by-trial basis, therefore we calculated the pupil response for each trial – which was defined as the pupil size in z-score minus the pre-cue baseline z-score (**Figure 4F**).

|  |
| --- |
|  |
| **Figure 4. Extracting the pupil information using DeepLabCut.**  (A) A schematic illustration of the pupillometry set up. The camera records the mouse’s right pupil while it is playing the matching pennies game. (B) An example frame showing both human labeling (dot) and DLC labeling (cross) for five labels, scale bar: 50 px. (C) X-axis offset of DLC labels from human labels (n = 324). The frames are taken from different sessions of different animals. (D) Y-axis offset of the same frames as in (C). (E) Top: An example trace of the z-score of pupil diameter in a single trial. Arrows indicate the time of the cue onset and the response. Bottom: average pupil traces for four different trial types: trials that the subjects chose left and got a reward; chose left and did not get a reward; chose right and got a reward; chose right and did not get a reward, from left to right. Gray shading indicates the standard deviation (F): An illustration that shows how the pupil response is calculated. The baseline is the average z-score of the -1 to 0 s before the cue of each trial. The pupil response at any time point (-3 to 5 s from the cue) is the baseline subtracted from the z-score at the corresponding time. |

**The pupil response contained choice- and outcome-related signals during matching pennies**

To characterize how the pupil responses were influenced by task-related variables, we used multiple linear regression. For each session, we fitted a regression model to determine the relation between the pupil responses and the choices, outcomes, reinforcers (choice-outcome interactions), the recent reward rate, and the cumulative reward (**Fig. 5A**). Specifically, the choices, outcomes, and interactions included terms for the next trial, the current trial, the last trial, and the trial before last to capture the potential persistent effects of these variables on neural correlates (Seo and Lee, 2007; Sul et al., 2010; Bari et al., 2019; Siniscalchi et al., 2019). The analyses revealed that pupil responses were modulated by choices, outcomes, and reinforcers during matching pennies (**Figs. 5B – D**). For a significant fraction of sessions, we detected a change in pupil size signaling the upcoming choice well before the next trial would start (cn+1**, Fig. 5B**). The early choice-related signal suggested that the animal was planning and preparing for the upcoming action prior to the cue. The choice-related signal ceased abruptly at the time of the cue onset for the next trial (cn-1**, Fig. 5B**). Meanwhile, outcome- and reinforcer-related signals in the pupil responses emerged after the potential reward would be delivered and persisted for a couple of trials (**Figs. 5C, D**). The pupil responses were also influenced by the recent reward rate and the cumulative reward (**Fig. 5E**). The cumulative reward was related to the number of trials performed, and therefore might reflect the motivational state of the animal. Although choice consistently influenced the pupil responses, the mean amplitude of the effect was more muted than outcome. Specifically, the presence of a reward led a large phasic dilation of the pupil, whereas the recent reward rate negatively modulated the pupil response (**Fig. 5F**).

|  |
| --- |
|  |
| **Figure 5. Pupil response is modulated by past reward information and upcoming choice.**  (A) A schematic diagram of the multilinear regression model that was fit to the pupil response in each 100 ms time bin. (B)The fraction of sessions with significant choice-dependent pupil activity against time, quantified by the linear regression model. The plots show the choice in the next trial (n+1), current trial (n), previous trial (n-1), and the trial before the previous trial (n-2), from left to right. The significance of each predictor was tested at α = 0.01. Red shading on the top indicates the p-value of the chi-square test. Colormap shows the p-value of the chi-square test. (C) Same as (B) for trial outcomes. (D) Same as (B) for the interactions of choice and outcome. (E)The fraction of sessions with significant coefficients of moving average reward in 20 trials (left), and the normalized cumulative reward up to current trial (right). (F) Regression coefficients of several predictors: choice of the current trial (*cn*), reward of the current trial(*rn*), average reward in a 20-trial running window (), normalized cumulative reward up to current trial (). Shaded areas are the 95% confidence interval estimated by bootstrap. |

**Pupil response was modulated by latent variables related to value updating, but not action selection**

Previously we showed that the animal’s behavior is well captured by the hybrid FQ\_RPE\_CK model, therefore we next asked whether there is influence of latent decision variables on pupil responses. To this end, we built additional multiple linear regression models using latent variables relevant for action selection (**Fig. 6**) or value updating (**Fig. 7**). For action selection, we considered the value difference, chosen value, choice kernel difference, and the chosen choice kernel (**Fig. 6A**). Consistent with our prior observation, we found significant choice- and outcome-related signals (**Fig. 6B - C**), however there was no reliable modulation of the pupil responses by the action-selection variables (**Fig. 6D - E**). For value updating, we tested variables including reward prediction error (RPE) and choice kernel error (CKE) in addition to value difference and choice kernel difference (**Fig. 7A**). In a significant fraction of sessions, pupil responses were modulated by the RPE, and to a lesser extent CKE and difference in choice kernel (**Fig. 7B – D**). Characteristically, positive RPE led to phasic dilation of the pupil, whereas negative RPE was associated with transient reduction in pupil size (**Fig. 7E**). Overall, these analyses show that transient change in pupil diameter is modulated by latent variables used in value updating, including RPE and potentially CKE.

|  |
| --- |
|  |
| **Figure 6. Pupil response is not modulated by variables related to choice selection.**  (A) A schematic diagram of the multilinear regression model. (B)The fraction of sessions with significant choice and reward-dependent pupil activity against time, quantified by the linear regression model. The plots show the choice (c), reward (r), and the interaction (c x r) in the current trial (n), from left to right. The significance of each predictor was tested at α = 0.01. Red shading on the top indicates the p-value of the chi-square test. Colormap shows the p-value of the chi-square test. (C) same as (B) in the previous trial (n-1). (D) The fraction of sessions with significant latent variables-dependent pupil activity against time. The plots show the difference in action values (), the action value of the chosen action (), and the difference in the choice kernel (), from left to right. (E) The fraction of sessions with significant latent variable and average reward-dependent pupil activity against time. The plot shows the choice kernel value of the chosen action (), moving average reward rate in 20 trials (Reward rate), and the normalized cumulative reward of the current trial (Cumulative reward), from left to right. |

|  |
| --- |
|  |
| **Figure 7. Pupil response is modulated by RPE.**  (A) A schematic diagram of the multilinear regression model. (B)The fraction of sessions with significant choice and reward-dependent pupil activity against time, quantified by the linear regression model. The plots show the choice (c) in the current trial, the choice and reward (r) in the previous trial (n-1), from left to right. The significance of each predictor was tested at α = 0.01. Red shading on the top indicates the p-value of the chi-square test. Colormap shows the p-value of the chi-square test. (C) The fraction of sessions with significant latent variable-dependent pupil activity against time. The plot shows the difference in action values (), the RPE, the difference in choice-autocorrelation factors (), from left to right. (D) The fraction of sessions with significant latent variable and average reward-dependent pupil activity against time. The plot shows the choice kernel error (), moving average reward rate in 20 trials, and the normalized cumulative reward of current trial, from left to right. (E) The regression coefficients for RPE and CKE. Shaded areas show the confidence interval estimated by bootstrap. |

**Performance and modeling of behavior in a two-armed bandit task**

The results presented so far indicated that for mice playing the matching pennies game, the transient pupil dilations were modulated by choice, outcome, and latent decision variables for value updating. To determine if the finding would generalize to a different task, we trained mice to perform a two-armed bandit task. Trials were organized into blocks, and each block was associated with one of two sets of reward probabilities (**Fig. 8A**). The criterion for a block switch was as follows: the subject must select the high-reward-probability side 10 times, and then subsequently the probability of a block switch happening in the next trial is 1/11. Therefore, the switch contained a random component, and the mouse could not predict the switch with certainty. Importantly, the timing of each trial is the same between the matching pennies game and the two-armed bandit task (**Fig. 1C**). **Figure 8B** presents an example session of a mouse performing the two-armed bandit task, showing reward probabilities reversing in blocks (top panel) and the mouse adapting its actions (bottom panel). The same hybrid FQ\_RPE\_CK model was used to quantify the behavior, with the estimated choice probability tracking the mouse’s choice pattern (**Fig. 8C**). Overall, mice performed the task well, averaging 702±26 trials per session with a 43.5±1.1% reward rate (n = 26 sessions from 4 mice; **Fig. 8D**). Model comparison based on BIC indicated that the hybrid model outperformed win-stay lose-switch and other Q-learning-based algorithms (**Fig. 8E**).

Although the hybrid model was successful in describing behaviors in both matching pennies and two-armed bandit task, there were notable differences in how the model operates for the two paradigms. Specifically, performance during two-armed bandit task involved weighted sums of action values and choice kernels – the key parameter for action selection – lying at the more extreme values (**Fig. 8F**, contrast with **Fig. 2D**). This difference was expected because in the two-armed bandit task, the animal spent considerable number of trials exploiting the high-reward-probability side in a block, and only needed to adapt around a block switch. Moreover, relative to matching pennies, the choice kernel had less weight (*β*K/(*β*K+*β*) = 0.29±0.05; **Fig. 8G**; p=0.02, two-sided Wilcoxon rank-sum test) and the inverse temperature sum was higher (*β*K+*β* =4.50±0.80; p=0.02), indicating that the animals relied more on reinforcement and had a lower tendency to explore. We again ran computer simulations to gain further insights into how the behavior might be affected by the balance of reinforcement learning and choice kernel (**Fig. 8H**). Unlike matching pennies, increasing the weight for choice kernels would lower the performance monotonically, and the animals appeared to lie away from the optimal strategy of using only reinforcement learning to guide decisions in this task.

|  |
| --- |
|  |
| **Figure 8. Mouse’s behavior in bandit task.**  (A) A schematic diagram of the bandit task. Left: Head-fixed mouse makes a left or right choice by licking the corresponding port. Right: The trials are separated into different blocks based on the high or low reward probability assigned to two ports. The exact reward probability is 0.7 and 0.1 for high and low, respectively. The block will switch randomly. (B) An example session of mouse performing bandit task. Top: the choices and outcomes. Black bar indicates a reward in the trial. The average reward rate is 52.8%. Blue vertical lines on top indicate right choices. Red vertical lines below indicate left choices. Bottom: the assigned reward probability of left (red) and right (blue) choices in different blocks. (C) An example of the time course of the latent variables and predicted behavior in the same session as (B). Top: Long vertical red and blue lines indicate rewarded left and right choices, respectively; short vertical red and blue lines indicate unrewarded left and right choices. Gray line shows the average probability to choose left, smoothed by a gaussian kernel. Black line shows the probability to choose left predicted by the model. Middle: the action values of left (red line) and right (blue line) choices estimated by the model; Bottom: the choice-autocorrelation factors of left (red) and right (blue) choices estimated by the model. (D) Summary of the behavior (26 sessions, 4 subjects). Top: the average trials performed every session is 702±132; Bottom: the average reward rate is 43.5%±5.7%. (E) Model comparison using BIC. The models were fit to different subjects, then the average of BICs over subjects were used to compare different models. (F) Psychometric curve fitted to the whole data set. Black histograms indicate the distribution of trials according to the weighted sum of the differences of action values (Δ*Q*) and the difference of choice-autocorrelation factors (Δ*K*). β, βK are the weights of action value and choice kernel, respectively. Dashed purple line shows the predicted probability to choose left according to the model; Purple dots show the probability to choose left computed from animal’s behavior. (G) Parameters estimation of the actual animal behavior. Left panel: the relative weight of choice kernel. *β*K/(*β*+*β*K) = 0.29±0.05; Middle panel: the total exploration level. *β*+*β*K = 4.50±0.80; Right panel: relative learning rate of the choice kernel. αK/α = 1.16±0.15. (H) The performance of a computational agent performing the dynamic foraging task. The agent adopted different values of *β*K/(*β*+*β*K), with the summation of inverse temperatures (*β*+*β*K) unchanged. The values of learning rate and summation of inverse temperature are set to be the median of the fitted value of the five animals. Black solid curve: the probability of stay as a function of different relative inverse temperature; Black dashed curve: average reward rate; Purple solid curve: entropy of three consecutive choices. Black solid, black empty, and purple dot: the probability of stay, average reward, and entropy of three consecutive choices of the computational agent whose all four parameters were set to be the median of the fitted parameters. |

**Similar pupil correlates for decision-related variables during two-armed bandit task**

We repeated the multiple linear regression analyses to identify factors that contribute to pupil responses during the two-armed bandit task. Similar to matching pennies, the choice- and outcome-dependent signals were present in a significant fraction of sessions (**Fig. 9A**).

The latent variables for action selection were largely absent (**Fig. 9B**). The exception is the chosen value, which was significant prior to cue, which should not have arisen prior to animal’s choice and may be due to the serial dependences in animal’s action during this task. The value-updating variables, especially the RPE and again to a lesser degree the CKE, modulated the transient pupil responses (**Fig. 9C**). Therefore, our results show that factors that influenced pupil responses in matching pennies – choice, outcome, and value-updating variables – were also contributors to the pupil responses during the two-armed bandit task.

|  |
| --- |
|  |
| **Figure 9. Multilinear regression of the decision variables in the dynamic foraging task.**  (A) The fraction of sessions with a significant choice- and outcome-dependent pupil activity against time, quantified by the linear regression model. The plots show the choice (top) and outcome (bottom) in the next trial (n+1), current trial (n), previous trial (n-1), from left to right. The results are taken from the same regression as equation (10). Red shading on the top indicates the p-value of the chi-square test. Colormap shows the p-value of the chi-square test. (B) Same as (A) for the latent variables. The plot shows the difference in action values (), the action value of the chosen action (), and the RPE, from left to right. The results are taken from the regression as equation (11). (C) same as (D) for the RPE and the choice kernel error (CKE), from left to right. The results are taken from the regression as equation (12). |

**Correlated fluctuations of left and right pupils during the two-armed bandit task**

One intriguing result is that pupil response was influenced by choice, i.e. whether the animal was choosing left or right. Is this a genuine choice-related signal? Some previous studies show that pupil dilations can predict the upcoming choice of human subjects including the choice timing and the selection of one out of five digits (Einhauser, Koch, & Carter, 2010), and the decision of a yes/no question (de Gee et al., 2014). However, another possibility is that when the animal makes a choice, the tongue lick movement could be associated with facial movements leading to spurious detection of pupil responses. To clarify the issue, we positioned two cameras to record both eyes simultaneously during the two-armed bandit task (**Fig. 10A**). For each eye, we applied multiple linear regression (**Fig. 5A**) to analyze the influences of choices and outcomes on pupil responses. We reasoned that if the choice-related signal was a movement artifact, then the aberrant signal should differ across eyes and across animals, and therefore the coefficients extracted from left and right eye would be uncorrelated. By contrast, if the choice-related signal was related to the animal’s internal decision, we would expect consistent dilation responses in both eyes. We analyzed the coefficients for the current choice between 3 – 5 s from cue onset when the pupil responses were largest (**Fig. 10B**). The choice-related signal for the left pupil were correlated with that for the right pupil in every session (r = 0.82, *P* = 3 x 10-40). The positive correlation coefficient indicated that the responses are symmetric in the two pupils. As a control, we plotted the coefficients of the current outcome, which was not lateralized and showed also a positive correlation coefficient as expected (r = 0.89, *P* = 0; **Fig. 10C**). Taken together, these results suggested that the effect of choice on transient pupil response could not be explained by simple movement artifact.

|  |
| --- |
| **Figure 10. Two-pupil recordings during dynamic foraging task.**  (A) schematics of the two-pupil recordings setup. Two cameras are placed in front of the two pupils with the same angle while the mouse is performing the task. (B) Scatter plot of the linear regression coefficients of the current choice within the 0-2 s from the cue time. The linear regression is the same as shown in Figure 4. X-axis: coefficients of the left pupil; Y-axis: coefficients of the right pupil. Different colors represent different subjects. Each dot represents a 0.1-s interval within the 0-2 s period. The black line shows the diagonal when coefficients of the left pupil equal to that of the right pupil. (C) Same as (B) for current outcome. |
|  |

**DISCUSSION**

The present study has two main findings. First, we demonstrated that head-fixed mice can play a two-player competitive game against a computer opponent. Their tendencies in the matching pennies game can be described in part by a computational model incorporating reinforcement learning and choice kernel. Second, we showed that transient pupil responses of the animals were associated with observable variables such as choices and outcomes, as well as latent variables for value updating, but not action selection.

**Performance in the matching pennies game**

Iterative matching pennies is a widely studied competitive game. Subjects playing the game tend to deviate from the Nash Equilibrium. For example, human players try to generate random choices but would switch too often (Camerer, 2003). Macaque monkeys and chimpanzees also showed deviation from the optimal strategy (Lee et al., 2004; Soltani et al., 2006; Seo et al., 2009; Seo and Lee, 2009; Martin et al., 2014). Rats were shown to counter-predict the opponent’s choice first and switch to a more random behavior when facing a strong competitor (Tervo et al., 2014). Moreover, pigeons playing the matching pennies game exhibit a similar divergence from the optimal play as the humans do (Sanabria and Thrailkill, 2009). Here we showed that head-fixed mouse can play matching pennies at a high level, albeit unsurprisingly also gaining rewards at below optimal rates. Demonstrating that mice can be studied using competitive game paradigms opens new avenues for studying neural circuitry for reward learning in animal models of neuropsychiatric disorders (Barthas et al., 2020; Liao and Kwan, 2021).

Although initial characterizations of the tendencies for sub-optimal play have relied on standard reinforcement learning algorithms(Erev and Roth, 1998b; Lee et al., 2004), recent studies have reported that deviations from the predictions of reinforcement learning (Hampton et al., 2008; Seo et al., 2014; Tervo et al., 2014). In this study, we employed a hybrid model, which combines the Q-learning with choice kernel, also known as the choice-autocorrelation factor (Katahira, 2015; Wilson and Collins, 2019). Specifically, the choice kernel was included to capture serial choice dependency, which is commonly observed in humans and animals performing various tasks (Akaishi et al., 2014; Abrahamyan et al., 2016). The hybrid model indeed fit the behavior better than variations of Q-learning algorithms (**Fig. 2**). We want to highlight the numerical simulations where the reward rate was examined as a function of inverse temperature (**Fig. 3**). The animal’s performance lied at a regime where further increase in reliance on choice kernel would deteriorate sharply the reward rate. The result hints at the possibility that the animal may be maximizing a reward-effort trade-off, because repeating the same choice is likely to be less effortful and indeed is the strategy taken by the animal often near the end of a session.

We further note that many studies of flexible decision-making in rodents have relied on two-armed bandit tasks with a block-based design (Ito and Doya, 2009; Sul et al., 2010; Sul et al., 2011; Tai et al., 2012; Bari et al., 2019; Groman et al., 2019; Hattori et al., 2019). This classic bandit paradigm has many merits, but also have a few shortcomings. First, the length of the blocks is a crucial parameter that relates to the volatility of the environment and is typically arbitrarily set by the experimenter. By contrast, matching pennies has a payoff matrix but otherwise no experimenter-defined parameters for the task. Second, within a block, it is advantageous for the animals to continually exploit the high-reward-probability option, therefore leading to strong serial choice dependencies. By contrast, in matching pennies, the computer opponent was designed to detect such dependencies and exert competitive pressure on the animal, therefore the animal is encouraged to always diversify its choice patterns during the session. Therefore, two-player games such as matching pennies are elegant and simple in design, and allows for investigation of neural mechanism underlying flexible decision-making under a regime that are quite different from two-armed bandit tasks (**Figs. 2D, 3, 8F-H**).

**Pupil responses and potential neuromodulatory mechanisms**

Pupil fluctuation is an indicator of the arousal state of an animal, and likely associates with the levels of various neuromodulator in the forebrain (McGinley et al., 2015). This relationship between pupil size and NE is supported by prior results, which showed reliable tracking of pupil fluctuations to activity of noradrenergic axons in the neocortex (Reimer et al., 2016). Furthermore, studies of the locus coeruleus (LC), the main source of NE for the forebrain, demonstrated a correlation between single unit firing in LC and pupil diameter (Aston-Jones and Cohen, 2005; Yang et al., 2021), and pupil dilation triggered by electrical microstimulation of the locus coeruleus (Joshi et al., 2016). Several studies have linked pupil dynamics and activity in LC to choice behavior, such as in the consolidation of the previous choices (Clayton et al., 2004; Einhauser et al., 2010) or the shaping of upcoming actions (de Gee et al., 2014). However, these studies were based on visual perceptual tasks, which is different from our task design where the cue is auditory and carries no relevant information except for trial timing. Our results hint at a potential role for noradrenergic signaling in post-decisional value updating, because the pupil response was correlated with the RPE and choice kernel error. This would agree with work that have showed the relevance for rewards in LC activity and pupil dilation in various behavioral settings (Sara and Segal, 1991; Aston-Jones et al., 1997; Varazzani et al., 2015). Furthermore, there is strong evidence linking pupil changes to errors and adaptations in decision tasks involving trial- or block-based inference (Nassar et al., 2012; Urai et al., 2017), highlighting the role of pupil-linked systems to control the influence of incoming data to guide future decisions.

We also showed that the pupil responses were modulated by the running reward rate and the total cumulative reward. These behavioral variables act on a longer timescale, therefore may reflect the noradrenergic regulation of long-term motivation. In the adaptive gain theory, LC neurons may exhibit phasic or tonic firing patterns, which reflect exploitation or exploration in behavior (Aston-Jones and Cohen, 2005). Our results show that the running average of the reward rate, which presumably is a major indicator for whether to continue exploiting or switch to exploring, modulates the pupil response, and therefore qualitatively is in accordance with the adaptive gain theory. The cumulative reward may be involved in the calculations of long-term utility, although more difficult to interpret because it also relates to other drifts across a session, such as the number of trials performed.

Overall, the current study lays the groundwork for studying reward-based learning in mice using competitive games. The two-player matching pennies game, which we showed the mouse can play against a computer opponent, may potentially be extended to mouse competing against another mouse in the future. The paradigm may therefore provide a quantitative framework for evaluating social decision-making in mice. The findings of significant pupil correlates to the major decision-related variables during matching pennies provide clues to the important roles of neuromodulatory mechanisms in mice. The current results open new avenues for future research into the role of neuromodulators in mediating the adaptive and social aspects of decision-making in mice.

**Acknowledgements**

We thank Daeyeol Lee and Hyojung Seo for helping with programming the matching pennies game, Michael Siniscalchi on behavioral training, and Matthew McGinley on the pupil recording setup. This work was supported by NIH/NIMH grants R01MH112750 (A.C.K.), R01MH121848 (A.C.K.), R21MH118596 (A.C.K.), China Scholarship Council-Yale World Scholars Fellowship (H.W.), Gruber Science Fellowship (H.O.), NIH training grant T32NS007224 (H.O.), and Kavli Institute for Neuroscience Postdoctoral Fellowship (H.A.).

**Author Contributions**

H.W. and A.C.K. designed the research. H.W. performed the matching pennies experiments. H.K.O. and C.E.M. performed the two-armed bandit experiments. H.W. analyzed the data; H.A. assisted by developing the bandit task and providing code for analysis. H.W. and A.C.K. wrote the paper with input from all other authors.

**References**

Abrahamyan A, Silva LL, Dakin SC, Carandini M, Gardner JL (2016) Adaptable history biases in human perceptual decisions. Proceedings of the National Academy of Sciences 113:E3548-E3557.

Akaishi R, Umeda K, Nagase A, Sakai K (2014) Autonomous mechanism of internal choice estimate underlies decision inertia. Neuron 81:195-206.

Aston-Jones G, Cohen JD (2005) An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Annu Rev Neurosci 28:403-450.

Aston-Jones G, Rajkowski J, Kubiak P (1997) Conditioned responses of monkey locus coeruleus neurons anticipate acquisition of discriminative behavior in a vigilance task. Neuroscience 80:697-715.

Bari BA, Grossman CD, Lubin EE, Rajagopalan AE, Cressy JI, Cohen JY (2019) Stable Representations of Decision Variables for Flexible Behavior. Neuron.

Barraclough DJ, Conroy ML, Lee D (2004) Prefrontal cortex and decision making in a mixed-strategy game. Nat Neurosci 7:404-410.

Barthas F, Hu MY, Siniscalchi MJ, Ali F, Mineur YS, Picciotto MR, Kwan AC (2020) Cumulative Effects of Social Stress on Reward-Guided Actions and Prefrontal Cortical Activity. Biol Psychiatry 88:541-553.

Bouret S, Sara SJ (2005) Network reset: a simplified overarching theory of locus coeruleus noradrenaline function. Trends Neurosci 28:574-582.

Camerer C (2003) Behavioral Game Theory: Experiments in Strategic Interaction: Princeton University Press.

Caze RD, van der Meer MA (2013) Adaptive properties of differential learning rates for positive and negative outcomes. Biol Cybern 107:711-719.

Chang SW, Gariepy JF, Platt ML (2013) Neuronal reference frames for social decisions in primate frontal cortex. Nat Neurosci 16:243-250.

Clayton EC, Rajkowski J, Cohen JD, Aston-Jones G (2004) Phasic activation of monkey locus ceruleus neurons by simple decisions in a forced-choice task. J Neurosci 24:9914-9920.

Dal Monte O, Chu CCJ, Fagan NA, Chang SWC (2020) Specialized medial prefrontal-amygdala coordination in other-regarding decision preference. Nat Neurosci 23:565-574.

de Gee JW, Knapen T, Donner TH (2014) Decision-related pupil dilation reflects upcoming choice and individual bias. Proc Natl Acad Sci U S A 111:E618-625.

Einhauser W, Koch C, Carter OL (2010) Pupil dilation betrays the timing of decisions. Front Hum Neurosci 4:18.

Erev I, Roth AE (1998a) Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. American Economic Review 88:848-881.

Erev I, Roth AE (1998b) Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. American economic review:848-881.

Gilzenrat MS, Nieuwenhuis S, Jepma M, Cohen JD (2010) Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. Cogn Affect Behav Neurosci 10:252-269.

Groman SM, Keistler C, Keip AJ, Hammarlund E, DiLeone RJ, Pittenger C, Lee D, Taylor JR (2019) Orbitofrontal Circuits Control Multiple Reinforcement-Learning Processes. Neuron.

Hampton AN, Bossaerts P, O'Doherty JP (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci U S A 105:6741-6746.

Haroush K, Williams ZM (2015) Neuronal prediction of opponent's behavior during cooperative social interchange in primates. Cell 160:1233-1245.

Hattori R, Danskin B, Babic Z, Mlynaryk N, Komiyama T (2019) Area-Specificity and Plasticity of History-Dependent Value Coding During Learning. Cell 177:1858-1872 e1815.

Hess EH, Polt JM (1960) Pupil size as related to interest value of visual stimuli. Science (New York, NY) 132:349-350.

Ito M, Doya K (2009) Validation of decision-making models and analysis of decision variables in the rat basal ganglia. J Neurosci 29:9861-9874.

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 23:1587-1596.

Joshi S, Li Y, Kalwani RM, Gold JI (2016) Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. Neuron 89:221-234.

Katahira K (2015) The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior. Journal of Mathematical Psychology 66:59-69.

Katahira K (2018) The statistical structures of reinforcement learning with asymmetric value updates. Journal of Mathematical Psychology 87:31-45.

Lee D (2008) Game theory and neural basis of social decision making. Nat Neurosci 11:404-409.

Lee D, Conroy ML, McGreevy BP, Barraclough DJ (2004) Reinforcement learning and decision making in monkeys during a competitive game. Brain Res Cogn Brain Res 22:45-58.

Liao C, Kwan AC (2021) Applying Reinforcement Learning to Rodent Stress Research. Chronic Stress (Thousand Oaks) 5:2470547020984732.

Martin CF, Bhui R, Bossaerts P, Matsuzawa T, Camerer C (2014) Chimpanzee choice rates in competitive games match equilibrium game theory predictions. Sci Rep 4:5182.

Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, Bethge M (2018) DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci 21:1281-1289.

McGinley MJ, Vinck M, Reimer J, Batista-Brito R, Zagha E, Cadwell CR, Tolias AS, Cardin JA, McCormick DA (2015) Waking state: rapid variations modulate neural and behavioral responses. Neuron 87:1143-1161.

Nassar MR, Rumsey KM, Wilson RC, Parikh K, Heasly B, Gold JI (2012) Rational regulation of learning dynamics by pupil-linked arousal systems. Nat Neurosci 15:1040-1046.

Nath T, Mathis A, Chen AC, Patel A, Bethge M, Mathis MW (2019) Using DeepLabCut for 3D markerless pose estimation across species and behaviors. Nat Protoc 14:2152-2176.

Pajkossy P, Szőllősi Á, Demeter G, Racsmány M (2017) Tonic noradrenergic activity modulates explorative behavior and attentional set shifting: Evidence from pupillometry and gaze pattern analysis. Psychophysiology 54:1839-1854.

Qiyuan J, Richer F, Wagoner BL, Beatty J (1985) The pupil and stimulus probability. Psychophysiology 22:530-534.

Reimer J, McGinley MJ, Liu Y, Rodenkirch C, Wang Q, McCormick DA, Tolias AS (2016) Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. Nat Commun 7:13289.

Sanabria F, Thrailkill E (2009) Pigeons (Columba livia) approach Nash equilibrium in experimental Matching Pennies competitions. J Exp Anal Behav 91:169-183.

Sara SJ, Segal M (1991) Plasticity of sensory responses of locus coeruleus neurons in the behaving rat: implications for cognition. Progress in brain research 88:571-585.

Seo H, Lee D (2007) Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. J Neurosci 27:8366-8377.

Seo H, Lee D (2009) Behavioral and neural changes after gains and losses of conditioned reinforcers. J Neurosci 29:3627-3641.

Seo H, Barraclough DJ, Lee D (2007) Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. Cereb Cortex 17 Suppl 1:i110-117.

Seo H, Barraclough DJ, Lee D (2009) Lateral intraparietal cortex and reinforcement learning during a mixed-strategy game. J Neurosci 29:7278-7289.

Seo H, Cai X, Donahue CH, Lee D (2014) Neural correlates of strategic reasoning during competitive games. Science (New York, NY) 346:340-343.

Siniscalchi MJ, Wang H, Kwan AC (2019) Enhanced Population Coding for Rewarded Choices in the Medial Frontal Cortex of the Mouse. Cereb Cortex 29:4090-4106.

Siniscalchi MJ, Phoumthipphavong V, Ali F, Lozano M, Kwan AC (2016) Fast and slow transitions in frontal ensemble activity during flexible sensorimotor behavior. Nat Neurosci 19:1234-1242.

Soltani A, Lee D, Wang XJ (2006) Neural mechanism for stochastic behaviour during a competitive game. Neural Netw 19:1075-1090.

Sul JH, Jo S, Lee D, Jung MW (2011) Role of rodent secondary motor cortex in value-based action selection. Nat Neurosci 14:1202-1208.

Sul JH, Kim H, Huh N, Lee D, Jung MW (2010) Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. Neuron 66:449-460.

Sutton RS, Barto AG (1998) Reinforcement learning: An introduction: MIT press Cambridge.

Tai LH, Lee AM, Benavidez N, Bonci A, Wilbrecht L (2012) Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. Nat Neurosci 15:1281-1289.

Tervo DG, Proskurin M, Manakov M, Kabra M, Vollmer A, Branson K, Karpova AY (2014) Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. Cell 159:21-32.

Urai AE, Braun A, Donner TH (2017) Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. Nat Commun 8:14637.

Van Slooten JC, Jahfari S, Knapen T, Theeuwes J (2018) How pupil responses track value-based decision-making during and after reinforcement learning. PLoS Comput Biol 14:e1006632.

Varazzani C, San-Galli A, Gilardeau S, Bouret S (2015) Noradrenaline and dopamine neurons in the reward/effort trade-off: a direct electrophysiological comparison in behaving monkeys. Journal of Neuroscience 35:7866-7877.

Vickery TJ, Chun MM, Lee D (2011) Ubiquity and specificity of reinforcement signals throughout the human brain. Neuron 72:166-177.

Wilson RC, Collins AG (2019) Ten simple rules for the computational modeling of behavioral data. Elife 8.

Wood RI, Kim JY, Li GR (2016) Cooperation in rats playing the iterated Prisoner's Dilemma game. Anim Behav 114:27-35.

Yang H, Bari BA, Cohen JY, O'Connor DH (2021) Locus coeruleus spiking differently correlates with S1 cortex activity and pupil diameter in a tactile detection task. Elife 10.