Probabilistic Learning in a Social Setting: Behavioral & Physiological Patterning

**Abstract**

A central question in current cognitive science is how people form mental representations of their own and others’ actions, and how this affects behavior. We explore this by studying how decision-makers are influenced by a partner's actions and outcomes. We develop the social multi-armed bandit (SMAB) task. The multi-armed bandit (MAB) is a probabilistic learning task. A participant is presented with multiple target-options (‘arms’) with different, unknown reward probabilities. In the SMAB, participants complete a block of MAB trials while turn-taking. In this paradigm, the reward probabilities are different for both participants, so that being influenced by the con-specific is unfavorable. We do a between subject analysis of participants who do the multiplayer SMAB, and participants who do a single player version of the paradigm. We record behavioral measures, including eye-tracked fixation patterns, to characterize how people converge on a response strategy. We also use pupillometry to examine pupillary responses to reward outcomes. We use electroencephalography (EEG) to measure Event Related Potentials (ERPs), specifically neurophysiological measures of event expectancy related to updating cognitive models. While we are still processing the data, several patterns have arisen in the behavior and pupillometry data. Participants use a more rewarding maximizing strategy in Game 2, suggesting that learning occurs in both paradigms. However, the scores in Game 2 are significantly higher only in the Multiplayer paradigm. This steeper learning curve suggests that in Game 1, participants encode their partner’s reward outcomes as their own. In the pupillometry data, we found an increase in dilation following no reward outcomes in the Multiplayer paradigm, suggesting increased attention and arousal as compared to in the Single Player. We also found increased dilation during periods of exploratory behavior.

**Introduction**

As social creatures, humans make most of our decisions within a social context, whether we are in the presence of other people, or informed by observation of or advice from our peers. Despite this, decision-making is often studied on individuals, or with social interaction simulated, losing the embodiment of engaging with a partner. In this study, we place dyads in a real-time social interaction while engaging in a decision-making task under uncertainty.

Decision-making is a high-level cognitive process that builds on perception, memory, and attention. Making a decision involves perceiving and evaluating alternatives, and choosing actions. The actor often receives feedback, informing their learning and future action-choices. A Stimulus prompts a Response, which results in an Outcome. Cognition evolved to allow animals to retain mental representations of these Stimulus-Response-Outcome contingencies (S-R-O) (Bland & Schaeffer, 2012). While decision-making is widely studied across disciplines, components and sub-process have not been synthesized into a single theory, and much remains to be understood about the neural basis of decision-making, in particular decision-making and learning in social environments.

The behavior paradigm we used in the study is a modification of the classic multi-armed bandit task. Multi-armed bandit tasks are an established paradigm for studying probabilistic decision-making (Kogler et al., 2016). These tasks present multiple choices ("arms") and each choice has a reward probability that is unknown to the decision-maker, akin to a slot machine at a casino. In an exploratory learning task, participants must choose actions to either explore the environment, or continue choosing an action (arm) that previously yielded a reward, exploitation. Importantly, exploration tends to reduce overall gains in the short term but can ultimately yield a better representation of S-R-O options in the environment (Adrian et al., 2019).

Typically following a period of exploration in which the S-R-O contingencies are learned, humans/animals converge on one of two strategies: maximizing and matching. In a maximizing strategy, decision-makers choose the most rewarding arm. In a matching strategy, decision-makers choose each arm proportionally to its reward probability. While maximizing is the optimal strategy for problems with static reward probabilities like the SMAB in this study, matching has been observed in humans and other animals (Adrian et al., 2019).

While decision-makers can use strategies that are some combination of the two strategies, they can also observe conspecifics using behavioral strategies that they have not used themselves (Cohen et al., 2007). Particularly in uncertain conditions, observing and encoding others’ actions can be an embodied representation of S-R-O contingencies. However, imitating others is not always beneficial. Findings from other human-subject cognitive tasks have shown that people represent other’s S-R-O contingencies even when it is unnecessary, or disadvantageous. In this study, we explore the extent to which human decision-makers are influenced by observation of con-specifics in a task, when it is unfavorable to be (Adrian et al., 2019).

Although many human decisions occur in a social context, only a few studies have included multiple decision-makers in social versions of prediction tasks such as the multi-armed bandit, and even these have not investigated effects of real-world social interaction or observation on decision-making. We do a between subject comparison of participants who do the Single Player and participants who do the Multiplayer versions of the task. We use recent developments in EEG and eye-tracking wireless technology that allow participants more ecologically valid mobility in the task. Both participants wear an EEG headset and eye-tracker, allowing us to gather behavior and physiological data from both participants in a hyperscan setup (Babiloni et al., 2006).

We use the eye-tracker to do an analysis of pupil response to reward outcomes in the task. We explore the psychosensory pupil response (PPR). Unlike pupillary dilation in response to light, the pupil light response, the PPR is in response to arousing stimulus, thought, or emotion. It is characterized by a brief rapid response about 0.5-1s after the stimulus and a slower response modulated by higher cognition that is about 1-2s following the stimulus. The function of the PPR Is unknown. There is a theory that it is high cognition modulating basic pupil dilation. When people are focused and calm, they may need visual acuity and constrict. When they are aroused, they may need visual sensitivity and dilate (Mathôt, 2018).

Another influential theory is known as adaptive-gain, and connects the PPR to exploration and exploitation, particularly through the locus coeruleus (LC). The LC is active when an organism is aroused. The LC is also involved in the pupil dilation pathway, along with other regions like the hypothalamus and the superior cervical ganglion, which projects to the spinal cord, innervating the iris dilator muscle. The theory states that exploration, in this case characterized by a state of distraction and switching between tasks, is associated with high, sustained LC activity, and large pupils. Exploitation, on the other hand, is characterized by engaging in a single task and is associated with intermediate and more phasic LC activity, as well as an intermediate pupil size. During exploration, the pupil is dilated and less response, which mirror LC activity (Mathôt, 2018). Based on these findings, we are studying whether or not the adaptive gain theory translates to explorative and exploitative behavior in the SMAB, hypothesizing that the pupil will be more dilated and perhaps less responsive during periods of exploration.

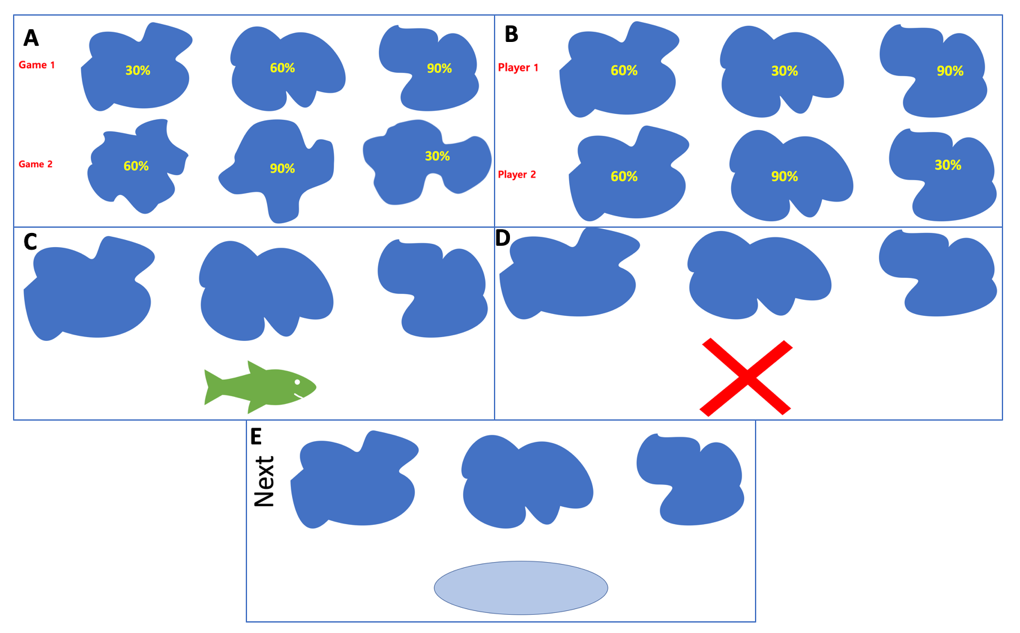
In addition to the adaptive gain theory, we are also looking for pupillary responses to reward outcomes. Pupil size fluctuations have been correlated with noradrenaline (NA) activity, which the LC is involved in projecting to the cortex. The dilation and NA-LC activity has also been correlated with uncertainty signaling. There is debate about pupil dilation modulation and pupillometric measurement. Evidence suggests that pupil dilation is greater under uncertainty and following negative feedback when expectancy is violated (Lavín *et al.,* 2014).

**Methods**

**Experimental Design**

The SMAB is a game displayed on a touchscreen, consisting of 100 trials per participant. Participants are told to earn as many points as possible by selecting one arm per trial, but will are not told anything else about the task. Our paradigm has three arms (i.e., locations) with reward probabilities of 30, 60, and 90. The ice hole shape and reward probability of each arm was randomized and different in Games 1 and 2. In the Multiplayer version, the arms with a probability of 30 and 90 are flipped for the participants.

In each trial, participants choose an arm and either get a reward or do not. A green fish signals a reward, while a red “x” signals no-reward. In the Single Player version, a confederate presses the “Next” button and a blue oval appears. We chose to include confederate trials in between subject trials in the Single Player paradigm in order to control for the effects of turn-taking on behavior, such as the slower pace of the task and the presence of a con-specific.



*Figure 1: Ice Fishing Task.* ***A****. Ice hole shapes are randomly assigned to a location, with different shapes being assigned to Game 1 and Game 2. Reward probabilities are randomly assigned to Game 1 and Game 2, with none of the probabilities being assigned to the same location in Game 2.* ***B****. The 30% and 90% holes are flipped for Player 1 and Player 2 in the Multiplayer version.* ***C****. When a subject gets a reward in a trial, the green fish appears, along with a positive reward sound.* ***D****. When a subject does not get a reward in a trial, the red “x” appears, along with a negative error sound.* ***E****. In the Single Player version, a confederate presses the Next button and a blue oval appears, in between the subjects’ trials.*

**Participants**

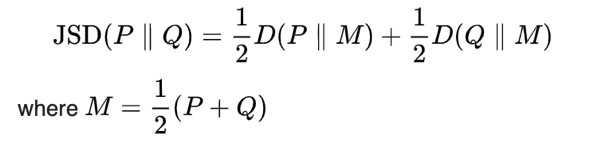
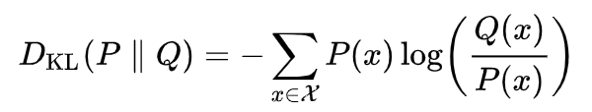
Participants in the Multiplayer version were 16 undergraduate students (4 males, 12 females). Participants in the Single Player version were 11 undergraduate students (6 males, 5 females). They were recruited through the university’s SONA system. They received course credit for participation, as well as a monetary reward based on performance in the task (0.05 USD per reward).

**Data Acquisition & Analysis**

The game was presented on a table-mounted touch screen monitor (diagonal: 66cm). The game was programmed in Python using the tkinter package. During the Multiplayer game, the participants sat facing each other. An Emotiv headset (www.emotiv.com) recorded 14-channel EEG data at a sampling rate of 128 Hz. A PupilLabs headset (pupil- labs.com/pupil/) captured eye-tracking and pupillometry data, with the eye camera sampling at 120 Hz.

PupilLabs Software was used to detect the pupil in each frame and calculate its diameter. Lab Streaming Layer (LSL) (Kothe, 2015) was used to synchronize all of the data streams (i.e. EEG, eye-gaze video, and game events) by time stamping each event and each sample. Synchronized EEG and pupillometry data were locked to participants' game choices in LSL-created XDF files so that behavioral and physiological data were epoched to trials.

On each turn, the 2 seconds of data following the outcome-stimulus presentation (win/loss) was used for further analysis. MatLab was used to export the XDF files and for pupillometry analysis. Python and the Pandas, Scipy, Numpy, and MatPlotLib packages were used for analysis of the behavior data.

 In the behavior data analysis, the first 20 trials were treated as training trials and omitted, except for in Figure 2. Participants’ strategies were analyzed using the Jensen Shannon divergence (JSD), which is based on the Kullback-Leibler divergence (KL). These are measures of relative entropy that can be used to compare relative distributions

Choice distributions in each game were compared to expected distributions based on probability learning behavior strategies. A maximizing strategy in this task is to choose the 90% arm 100% of the time. A matching strategy is to choose each arm in proportion to its reward outcome.

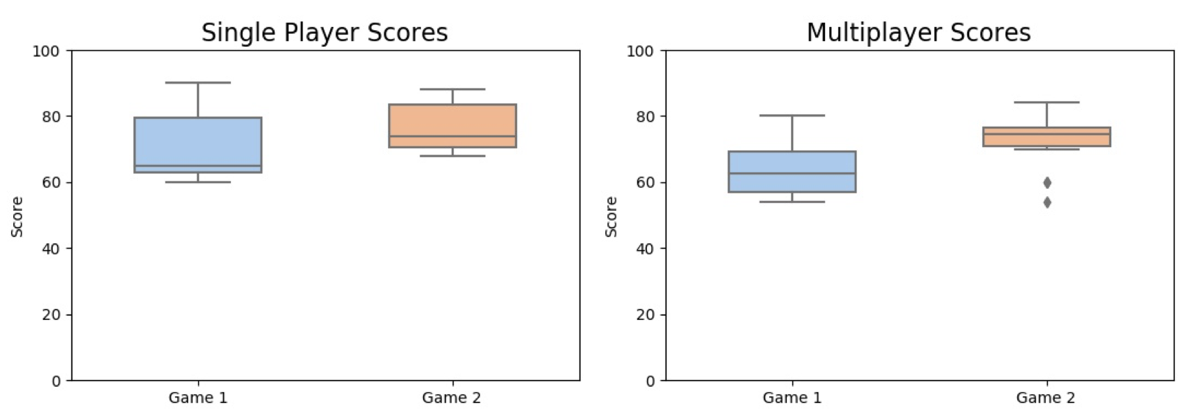
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Choice Distributions | Strategy | 30% Arm | 60% Arm | 90% Arm |
| Maximizing | 0% | 0% | 100% |
| Matching | 16.7% | 33.3 | 50% |

*Table 1. Expected choice distributions for the strategies.*

**Results**

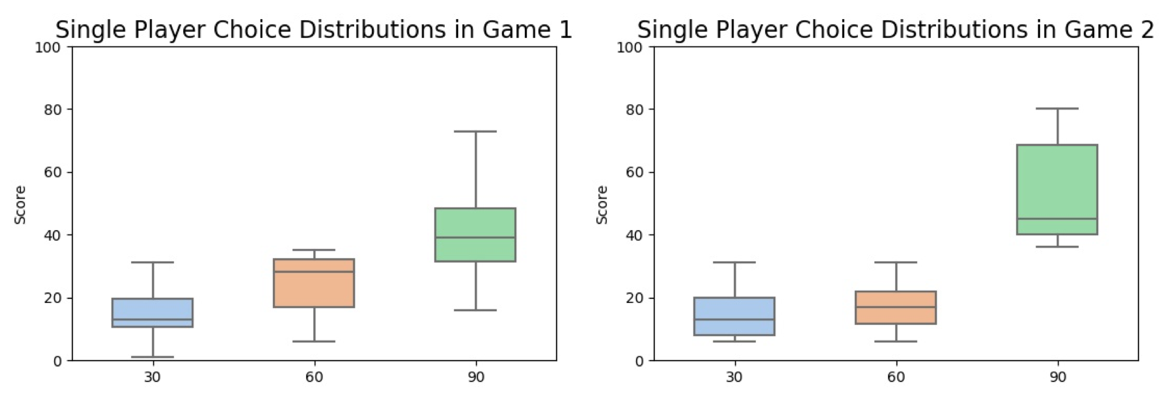
**Behavior**

In each trial of the game, participants were presented with outcomes after making a response. They learned that different arms were associated with different probabilities of receiving a reward, and used this information to make future decisions. We analyze participants’ scores and choice distributions in each game.

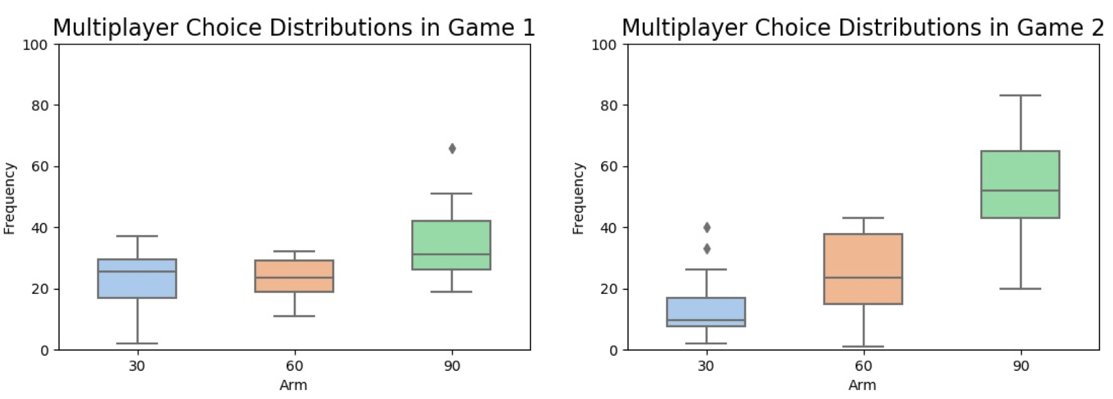


*Figure 2. Participant scores in the Single Player and Multiplayer paradigms. Single Player (left): difference between Games 1 and 2 n.s. (p = 0.108, n = 11). Multiplayer (right), improvement from Game 1 to Game 2 was significant (p = 0.004, n = 16).*

While the mean score in the Single Player paradigm is higher than in the Multiplayer for Game 1 and Game 2, this is not statistically significant. This may be due to the small sample size, especially in the Single Player paradigm. In the Single Player, the difference in the score between Games 1 and 2 is not significant (Figure 2). However, in the Multiplayer, the improvement from Game 1 to Game 2 was significant (Figure 2). This may mean that there was a steeper learning curve in participants learning not to encode their partners’ reward outcomes as their own.

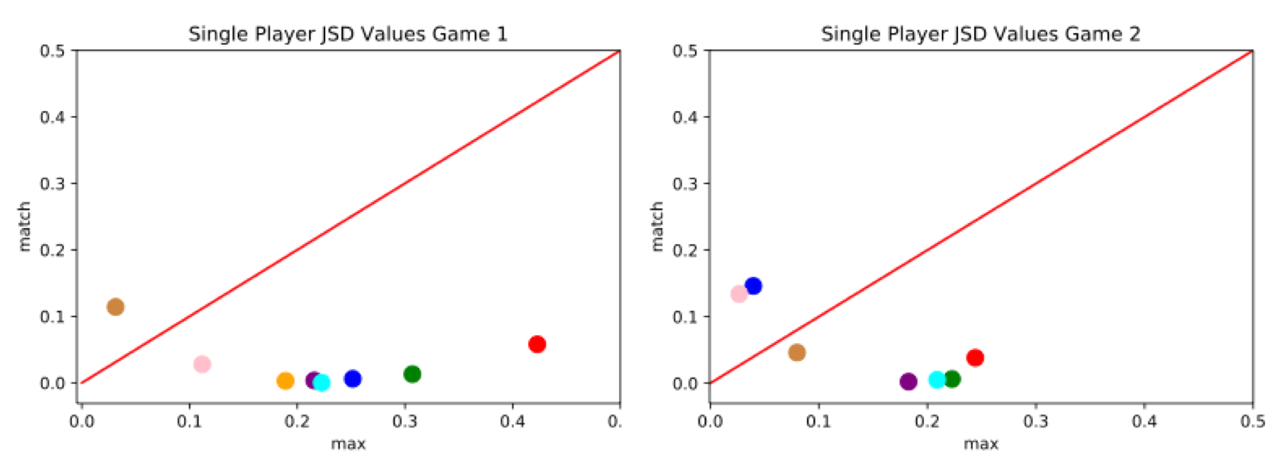


*Figure 3. Choice distributions in the Single Player paradigm, in Game 1 and Game 2 (n = 11). Hole 30 is chosen significantly less than hole 90 in both games (p = 0.0462 and 0.00694), while hole 60 is chosen significantly less only in Game 2 (p = 0.170 and 0.00910). The change in the number of times that hole 30 and hole 60 are chosen between Game 1 and Game 2 is n.s. (p = 0.147 and 0.122).*



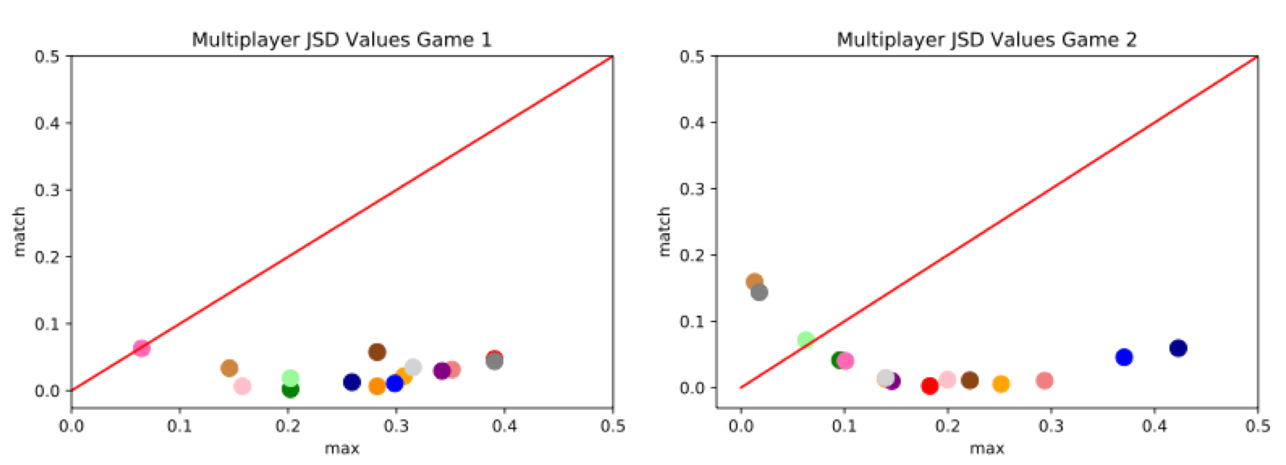
*Figure 4. Choice distributions in the Multiplayer paradigm, in Game 1 and Game 2 (n = 16). Hole 60 is chosen significantly less than hole 90 in both games (p = 0.0267 and 0.00145), while hole 30 is chosen significantly less only in Game 2 (p = 0.0622 and 5.670e-05). The change in the number of times that hole 30 is chosen between Game 1 and 2 is significant (p = 0.0198). The change in the number of times that hole 60 is chosen is n.s. (p = 0.910).*

The choice distribution data also reflects impaired learning, suggesting that participants encode their partners’ reward outcomes similarly to their own. In the Multiplayer, between Games 1 and 2, the change in number of times hole 30 is chosen is significant, while the number of times the hole 60 is chosen is not significant (Figure 4). Hole 30 for a participant is hole 90 for their partner. In Game 1, participants choose the 30 hole more than the 60 hole, but less than the 90 hole. This suggests that participants encode their partner’s reward outcomes, but less than their own. The lower frequency of hole 30 choices in the second game results in the significantly higher score in Game 2. However, the 60% hole may still be chosen in Game 2 because it collectively yields the most rewards for both partners combined. Choosing the 60% hole may be a cooperative decision.



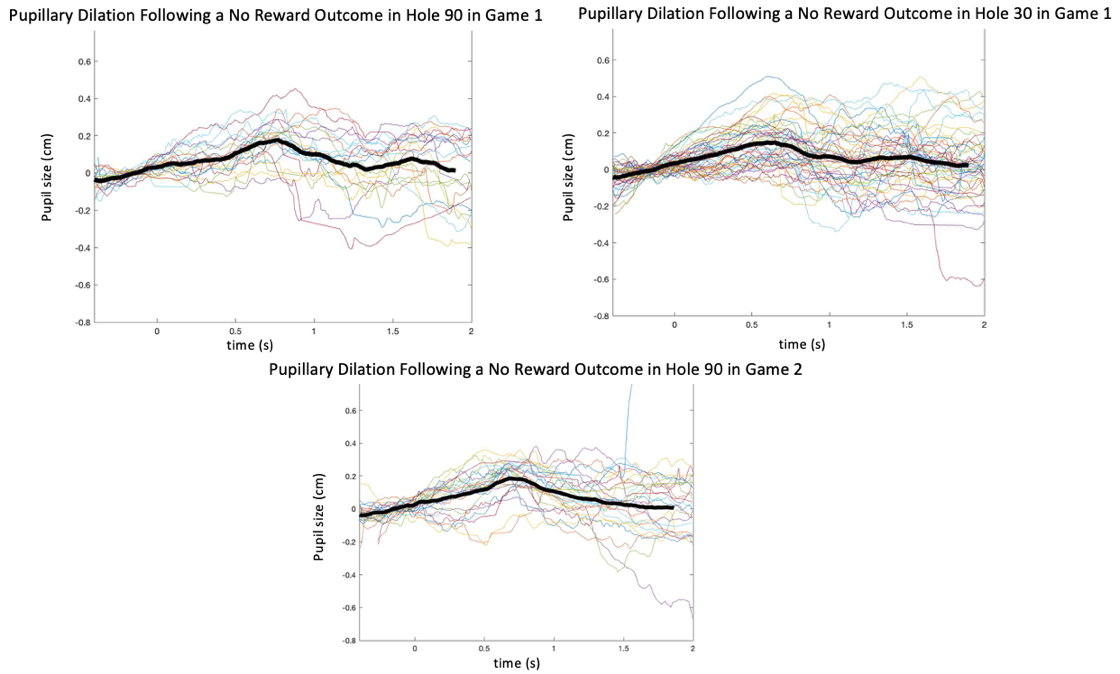
*Figure 5. JSD values for predicted maximizing and matching responses in the Single Player paradigm, in Game 1 (left) and Game 2 (right). Each dot corresponds to a subject. The red slope line indicates an idealized division between preference for one strategy or the other: subjects under the red line had choice distributions that favored a matching strategy, whereas subjects*

*above the line had choice distributions that favored a maximizing strategy. Between Game 1 and Game 2, JSD values shifted significantly more towards maximizing choice (p = 0.0608, n = 11).*



*Figure 6. JSD values for predicted maximizing and matching responses in the Single Player paradigm, in Game 1 (left) and Game 2 (right). Each dot corresponds to a subject. Subjects under the red line had choice distributions that favored a matching strategy, whereas subjects above the line had choice distributions that favored a maximizing strategy. Between Game 1 and Game 2, JSD values shifted significantly more towards maximizing choice (p value = 0.0160, n = 11).*

Next, we analyzed the choice distributions by calculating the JSD between the distributions and idealized distributions based on a matching and maximizing strategy, we found that most of the participants favored a matching strategy. For both the Single Player and Multiplayer versions, participants’ choices shifted more towards a maximizing strategy.



*Figure 7. Pupillary dilation following a no Reward outcome, in hole 90 in game 1 (left) and hole 30 (right). Each line corresponds to a trial and the black line is the average. The data is separated into trials that begin 0.5 s and end 2 s after the onset of a reward. The data has been filtered to eliminate diameters outside of the normal range, as well as to remove blinks. The peak following a no reward outcome in hole 90 is greater, 0.176 versus 0.148, and there is a peak at about 1.6 s that is absent in hole 30.*

While we are still analyzing the pupillometry data, preliminary analyses suggest an increased dilation, as well as a second dilation following a no reward outcome in hole 90, as compared to a no reward outcome in hole 30.

**Discussion & Further Research**

We further developed the Social Multi-Armed Bandit task to examine the influence of social interaction on decision-making. Participants’ choice distributions converge on a more rewarding maximizing strategy in Game 2, suggesting that learning occurs in both paradigms. However, the scores in Game 2 are significantly higher only in the Multiplayer paradigm. This steeper learning curve suggests that in Game 1, participants encode their partner’s reward outcomes as their own. This pattern is also observed in the choice distributions. In Game 1, participants choose hole 30, which has the lowest reward probability for them and the highest reward probability for their partner, more than hole 60, which yields the same reward probability for both partners. In Game 2, they choose hole 30 significantly less but choose hole 60 at a similar rate. This suggests that participants encode their partners’ reward outcomes as their own in Game 1, and less so in Game 2. However, they do may continue to choose hole 60 because it yields an intermediate reward for both participants, making it a cooperative choice.

Preliminary pupillometry data suggests that participants have a greater pupillary dilation following a no reward outcome in hole 90. This is consistent with other studies, which have found a pupil dilation following negative feedback that violates expectations. As hole 90 yields a reward 90% of the time, a no reward outcome from this hole is unexpected (Lavín et al., 2014). In our study, feedback is used to inform future decisions. The violation of expectation may result in an updating of the S-R-O contingencies of the task. The pupil dilation may be a marker of NA signaling, which is involved in learning under uncertainty (Lavín et al., 2014). When participants are not rewarded by a response that rewards them 90% of the time, their expectation may be violated, activating NA to update uncertainty and improve learning, also causing pupillary dilation. Thus, surprise may play a role in behavior adaptation under uncertainty (Yu and Dayan, 2005).

However, this data is preliminary and requires further analysis. An analysis of trials in which participants are using more exploratory and exploitative behavior in comparison to pupillary dilation across those trials may yield insight into the adaptive-gain theory. The EEG and eye-tracking data, not reported here, also require further analysis. A further analysis of the behavior data may reveal individual differences in game strategies, which may be correlated with differences in the physiological data.

**Acknowledgements**

**References**

Bland A, Schaefer  A. Different Varieties of Uncertainty in Human Decision-Making. Frontiers in Neuroscience. 2012;6:85. doi:10.3389/fnins.2012.00085

Kogler L, Sailer U, Derntl B, Pfabigan DM. Processing expected and unexpected uncertainty is modulated by fearless-dominance personality traits – An exploratory ERP study on feedback processing. Physiology & Behavior. 2017;168:74-83. doi:10.1016/j.physbeh.2016.10.016