# 

# 

# 

# 

# 

# 

# 

# Exploring Model-Based Vs Model-Free Pupillometry Correlates to Reinforcement Learning Parameters

Chris Endemann

Computational Cognitive Science

# Introduction

Reinforcement learning (RL) algorithms are particularly unique insofar as they are arguably the most successful in their ability to develop a model of the world without explicit supervision. The luxury of label evasion is afforded to reinforcement models through their strategy of evaluating the world and the agent’s actions through value-based lenses. For instance, in Mnih et al. (2015)’s deep Q-network model, a feature map is developed within a deep artificial neural network architecture which receives feedback based on prediction-errors in value judgments of a given action in a given state. Given that our environments do not often provide for explicit labels for supervised learning, and pure unsupervised learning (i.e. learning based solely off of statistical regularities of a veridical feature representation of the external world) tends to lead to a high number of possible solutions to a single classification problem (with few available verification procedures), a value-based approach towards sensory model-updating is a fairly appealing idea for explaining the brain’s ability to efficiently and effectively represent our external environments in an effort to guide behavior.

In further support of these ideas, there is a growing number of studies which have successfully used reinforcement learning models to capture and explain both behavioral and neurobiological variables and processes during human decision making tasks. One now famous example of such evidence includes modeling of the dopaminergic system in response to reward, with dopamine neurons producing tell-tale activation profiles reflecting the temporal difference algorithm (Waelti et al., 2001). In more recent years, many studies have examined and tested variants of reinforcement learning models to better predict decision making and learning processes, as well as to find variables that explain the mechanisms behind such processes (e.g. Gläscher et al., 2010; Houillon et al., 2013; Jepma & Nieuwenhuis, 2011; Lee et al., 2012). In common to all of these studies is their ability to either propose more explanatory/accurate models of decision-making processes through model-comparison tests--sometimes referred to as the model learning approach--or to regress externally measured random variables against fitted parameters of the reinforcement learning models to infer relationships between model variables and other external measurements--sometimes referred to as the observation model approach (Daw, 2009).

In the current study, the observation model approach is taken in order to emphasize a predicted dynamic interplay of pupil constriction in response to the recollection of sensory-state information as well as pupil dilation in response to a directed focus of bottom-up processing (i.e. an increase in the uptake and/or reliability of sensory information) for sensory-model updating. While pupillary responses have been used in several papers to-date in order to examine the role and impact of both gaze direction and pupil dilation on decision making and attentional processes (Koenig et al., 2017; Lavín et al., 2014; Preuschoff et al., 2011), such studies have predominantly framed the dilation of the pupil as an indicator of risk, certainty or surprise, and attention. The present paper side-steps the more frequently assumed relationships between pupil size and decision-making processes, and instead frames the dynamics of the pupil as an indicator reflecting the interplay between bottom-up (referring to information traveling from the external environment to an internal representation) sensory model updating and top-down processing specifically involving the recollection/consolidation of past information and simulation of future behaviors/decisions.

This framework has a wide range of behavioral and neurological evidence in support of it. First off, in terms of modeling studies, it has already been shown that pupil dilation tends to be increased preceding exploratory behavior defined in RL terms (i.e. behavior which ignores the presently optimal and familiar action in a given state in hopes of encountering a longer-term reward that outperforms the agent’s current and familiar approach), and participants with larger pupil diameters, on average, tend to engage in more exploratory behavior (Jepma & Nieuwenhuis, 2011). If one assumes that exploratory behavior over the environment is likely to simultaneously foster an increased propensity for bottom-up sensory model-updating rather than predictive processing based on sensory expectations--an assumption that seems appropriate given 1) the fact that exploration takes place in novel/unfamiliar situations, and 2) the amount of evidence suggesting that the brain tends to update sensory-models when faced with novel situations or stimuli (Friston, 2010; Spratling, 2017)---then one can reasonably interpret these findings as suggesting that increased pupil dilation corresponds to the brain increasing its gain towards bottom-up information streams in hopes of updating internal sensory models in a manner that optimizes behavior. As additional support for this interpretation, studies relating pupillometry measures to indicators of surprise find that the pupil dilates in response to the absence of an expected stimulus as well as to the stimulus, itself (Levine, 1969).

Furthermore, when sensory models are not being updated during a decision making task, the present theory proposes that the brain is then more likely to be engaged in extracting previously stored sensory model information for exploitative purposes and/or consolidating past information sequences. Relating this concept to the physiology of the pupillary response, it has been shown that in rats, pupil constriction strongly correlates with the rate of sharp wave ripples (SWRs) stemming from the hippocampus (McGinley et al., 2015). SWRs are implicated in memory retrieval and consolidation processes, including the retrieval/re-activation of spatial information via place cell sequence recreation during rest or quiet wakefulness (Atherton et al., 2015; Foster & Wilson, 2006). In addition, more recently, SWRs have also been implicated in the generation of simulated preplay of future decisions (Buhry et al., 2011; Buzsáki et al., 2015)--meaning that SWRs and pupil constriction might also indicate planning behavior via the simulation of future events. Therefore, and assuming the rodent model applies to humans in this case, the aforementioned correlation allows one to relate pupil constriction to the likelihood that one is to be engaged in either memory retrieval, memory consolidation, or event preplay processes.

Taking these findings together, it would seem that the pupil can indeed reflect the particular processing mode of the brain--specifically--whether the brain is engaged with an information encoding process or an information extraction process. Interestingly, such a dichotomous view of the brain’s processes is also supported by studies which examine the circuitry utilized to unfold sensory recollections. For example, if one engages in visual imagination of a given image, the fMRI bold response is strikingly similar throughout both recollection and sensation processes (Ganis et al., 2004). Furthermore, recall processes can even lead to contamination of sensory representations if one is simultaneously engaged in memory retrieval during sensation (Kang et al., 2011); such findings may partially explain why SRWs are typically seen only during quiet wakefulness or sleep (Foster and Wilson, 2006). Thus, given this observed interplay between these two processes, it is not entirely far-fetched to expect a physiological correlate to bottom-up processing, such as the pupil response, which inverts its response during the recollection of information (thereby reflecting inhibition of sensory processing and increased memory processing as to not confuse the direction of information flow within the cortex).

To study and foster support for this framework in an RL context, a hybrid Q-learning model that takes into account both model-free and model-based learning strategies is employed to fit human decision-making data--generously provided by Konovalov and Krabjich (2015)--to recover measurements of exploratory behavior (allowing for an indirect analysis of the relationship between pupil size and sensory model-updating) and exploitative behavior (allowing for an indirect analysis of the relationship between pupil size and memory-recollection processes). The use of a hybrid model has several advantages in terms of the model’s relatability to this paper’s theorized dynamics of the pupillary response.

For one, the model allows one to classify participants as taking a predominantly model-free approach towards task learning or a model-based approach. Model-free learning, in many respects, can be thought of as more of a habit-oriented, unconscious association behavior. The error signal guiding behavior in a model-free RL model simply amounts to the difference between the actual and expected reward in a given state. No further modeling of the environment is needed aside from an evaluation of the current state based off of prior experience of the gains associated with that state. A model-based approach, on the other hand, involves learning the structure of the task in order to form predictions based off of the current state and the next likely transition. Actions can then be evaluated by searching this model.

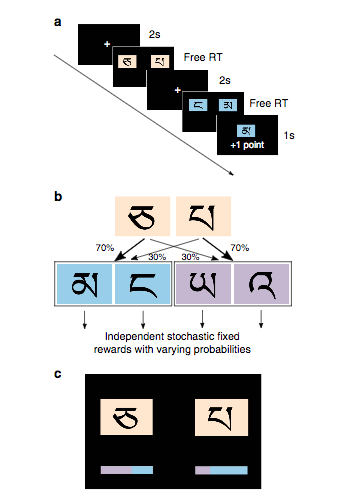
In terms of the pupillary response, it is necessary to classify learning strategies as either model-free or model-based given the different expectations for each subgroup. Model-free learning--given the lack of an adequate model of the task structure--is predicted to lack as much engagement with recall processes (given the environmental information is assumed to be unconsciously associated and learned). This reduction in recall processes is hypothesized to manifest in the data as an increased average pupil dilation during exploitative decisions amongst model-free learners as compared to model-based learners. Regarding bottom-up sensory model-updating processes, however, discrepancies between learning strategies and pupillometry measurements seem less evident. That is, both learners must represent their environments within the sensory cortices, and there doesn’t seem to be any obvious reason to suspect model-based learners to have a higher gain of sensory information simply due to the fact that they are building a model of the task rather than solely a model of the sensory environment (which both learners must do). Therefore, it is expected that pupil dilation levels will be similar between the two subgroups during exploration trials.

By investigating these predictions and the framework that supports them, the hope is not only to foster support for the theory laid out in this paper, but also to find reliable and easily measurable behavioral variables (e.g. the pupillary response) that can help better predict decision making processes and better explain the dynamics of cortical function. That is, there is still a good deal of unaccounted variance in human decision making datasets fitted to RL models arguably due to the too-frequent exclusion of two vital components of human learning within the model frameworks--the reliability and dynamics of retrieving stored information (Q-values in an RL framework)--though see Collins and Frank, (2012)for an exception**--**and the likely complicated manner in which value-estimates are actually updated during learning.

# Methodology

## Dataset/Behavioral Task

The data used to test these models was borrowed from Konovalov & Krabjich--*Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning* (2015)-- who utilized a two-stage markov decision task originally designed by Daw et al. (2011) to probe the effects of gaze direction in both model-based and model-free learning strategies. Their study design is recapitulated here. Forty-three subjects completed the experiment, which consisted of two conditions with 150 trials each (though only the first condition is analysed in this paper due to luminance covariates introduced in the second condition). In the first stage of the task, subjects had to make a choice between two Tibetan symbols that could lead to one of two second-stage states, ‘purple’ and ‘blue’ (Fig. 1a). The transition was stochastic: one symbol chosen in the first stage was more likely to lead to the blue state, and the other one was more likely to lead to the purple state. Thus each first-stage symbol had a ‘common’ state (probability 0.7) and a ‘rare’ state (probability 0.3) associated with it. Once one of the states was reached, subjects had another choice between two additional symbols of the respective colour (Fig. 1b). Each of the four second-stage symbols had an independent probability of yielding a fixed reward. During the course of the experiment, these probabilities drifted independently in the range from 0.25 to 0.75 according to slow Gaussian walks with mean=0 and s.d.=0.025. Drifting probabilities were added with the intent of facilitating active learning and exploring different states. In each stage, the position of the symbols on the screen was randomized. Choices were made using a keyboard, and every choice was followed by a white frame around the chosen symbol for 0.5 s. All choices had free RT. After the second-stage symbol was chosen, it was displayed at the center of the screen, with the outcome shown in the bottom part of the screen (either ‘+1 point’ or ‘0 points’). Before starting the task, subjects were introduced to the rules of the task, including a short practice on each part of the task, and a 30-trial practice session with different stimuli. In addition, every 50 trials, participants were allowed a brief break.



**Figure 1. Task Design**

*A representation of the two-stage decision task. Panel a. exemplifies the time-course of a typical trial (rewarded in this case). Panel b. outlines the organization of transitions between states. The tan Tibetan symbols (stage-1) lead to either the blue pair of stimuli, or the purple pair (collectively referred to as stage-2 stimuli), with one pair dominating the transition from the first chosen symbol. Panel c. gives details regarding condition two of the behavioral task, which was not explored in this paper (figure borrowed from Konovalov & Krabjich, 2015).*

## 

## Hybrid Model

The model used for fitting decision making data to an RL framework was mimicked from Konovalov and Krajbich (2011). These authors used a hybrid learning model which combines model-free and model-based components along with a parameter to control for how strongly each learning strategy plays into evaluations of state-action Q-valuesnotated by *w*.

The model-free learning strategy uses only reward information to update the Q-values. These values are initialized to zero at the beginning of the each trial-set. Let *a*1 be the symbol chosen in the first stage of the task, and *a*2 be the second-stage choice (subscripts indicate stage number). Then, after a trial *t* is completed and a reward *r*(*t*)∈{0,1} is received, the chosen second-stage symbol *a*2’s Q-value, Q2, is updated in the following manner:



where α is a learning rate parameter, and *r*(*t*)−Q2MF(*i,t*) is a reward prediction error.

The value of the symbol chosen in the first stage is also updated through the reinforcement process, using both the second-stage reward prediction error and the prediction error that comes from the difference between the obtained and expected value of the second-stage state:



where λ is an eligibility trace parameter that captures the effect of the second-stage prediction error on the first-stage action value.

The model-based learning strategy incorporates the empirical transition probabilities into the updating process (Otto et al., 2013):



where P(blue|*a*1) and P(purple|*a*1) are the respective transition probabilities after choosing action *a*1 which are calculated using Beta-Binomial Bayesian updating:



where N(blue|*a*) and N(purple|*a*) are the numbers of times the blue or purple state was reached after making a choice *a*.The hybrid model Q-value for each first-stage choice is calculated using a combination of the model-free and model-based action values:



where *w* is a weight parameter restricted between 0 (pure model-free strategy) and 1 (pure model-based strategy).

In order to output predicted decisions, a logit discrete choice model is assumed for both stage choices, with probability of the second stage choice computed as



## 

and the first stage as



where *β* is the ‘inverse temperature’ parameter that controls for how likely the agent is to follow suit with its Q-value estimations or choose a more exploratory option. The *p* parameter in the equation above represents the stickiness of a given choice in a trial-set. This parameter was removed from this paper’s analysis given its lack of ability to produce sufficient additional explanatory power (reduce BIC values).

## 

## Model Fitting Procedure

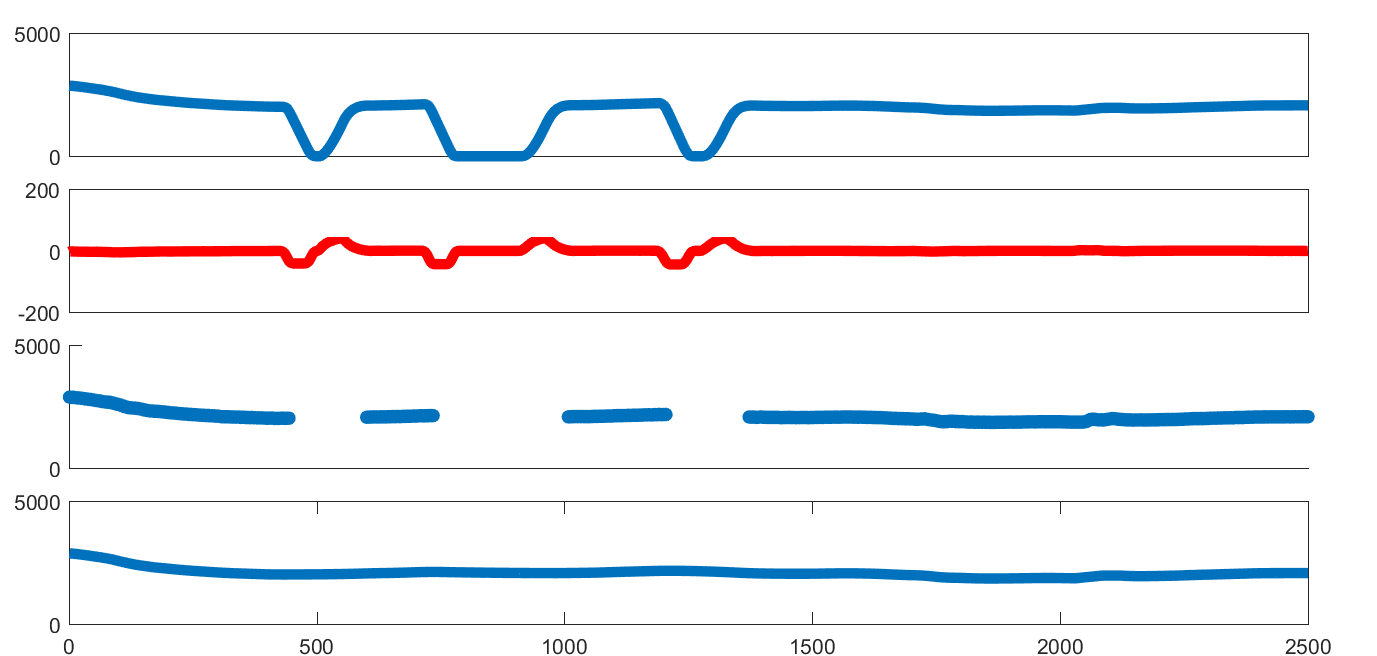
The model was fit individually for each subject’s data using a maximum likelihood estimator and the probability formulas defined above. Parameters α, λ and w were restricted to lie between 0 and 1, and *β* between 0 and 10. To achieve a global maxima of the estimator function and recover the most likely parameter fits for each participant’s data, Matlab’s *multmin* function was utilized with 50 random restarts. Predicted (hybrid) Q-values for each trial, state, and choice, are then extracted from the fitted parameters for further analysis.

## Eye-tracking methods

Verbatim from Konovalov and Krajbich (2011): Subjects’ gaze data was recorded using an EyeLink 1000+ desktop-mounted eye-tracker with a chin rest and sampled at 1000 Hz. Before every choice, subjects were required to fixate at the center of the screen for 2 s, or the software did not allow them to proceed. This ensured unbiased initial gaze positions. The task was created and displayed using Matlab and Psychtoolbox59. The chin rest was placed at 65 cm away from the screen, and the screen resolution was set at 1920 × 1080.

## Eye-tracking analysis

Before using pupil-size as a measured variable, it was necessary to clean the pupillometry data. When measuring pupil size, the standardized procedure typically involves first smoothing the raw response to ignore small fluctuations (Fig. 2A), followed by detection of blinks through a velocity threshold parameter indicating the speed at which the eyes are opened or closed (Fig. 2B). Removed blinks (Fig. 2C) are replaced via interpolation (Fig. 2D). This standard procedure was followed, and missing points were interpolated using Matlab’s *interpl* function--with the ‘*pchip*’ technique/flag yielding consistent results and satisfactory replacement values. In accordance with Jepma and Nieuwenhuis (2011), a nearly identical period of interest (500 ms longer) was used for measuring pre-choice pupil responses. That is, average pupil dilation for a given trial was measured from 3000 ms preceding a decision (keyboard press) to 500 ms preceding a decision--factoring out the influence of the commonly observed anticipatory pupillary response which immediately precedes a decision. Each participant’s average pupil diameter for each trial are plotted in Figure 3 for reference.



A.

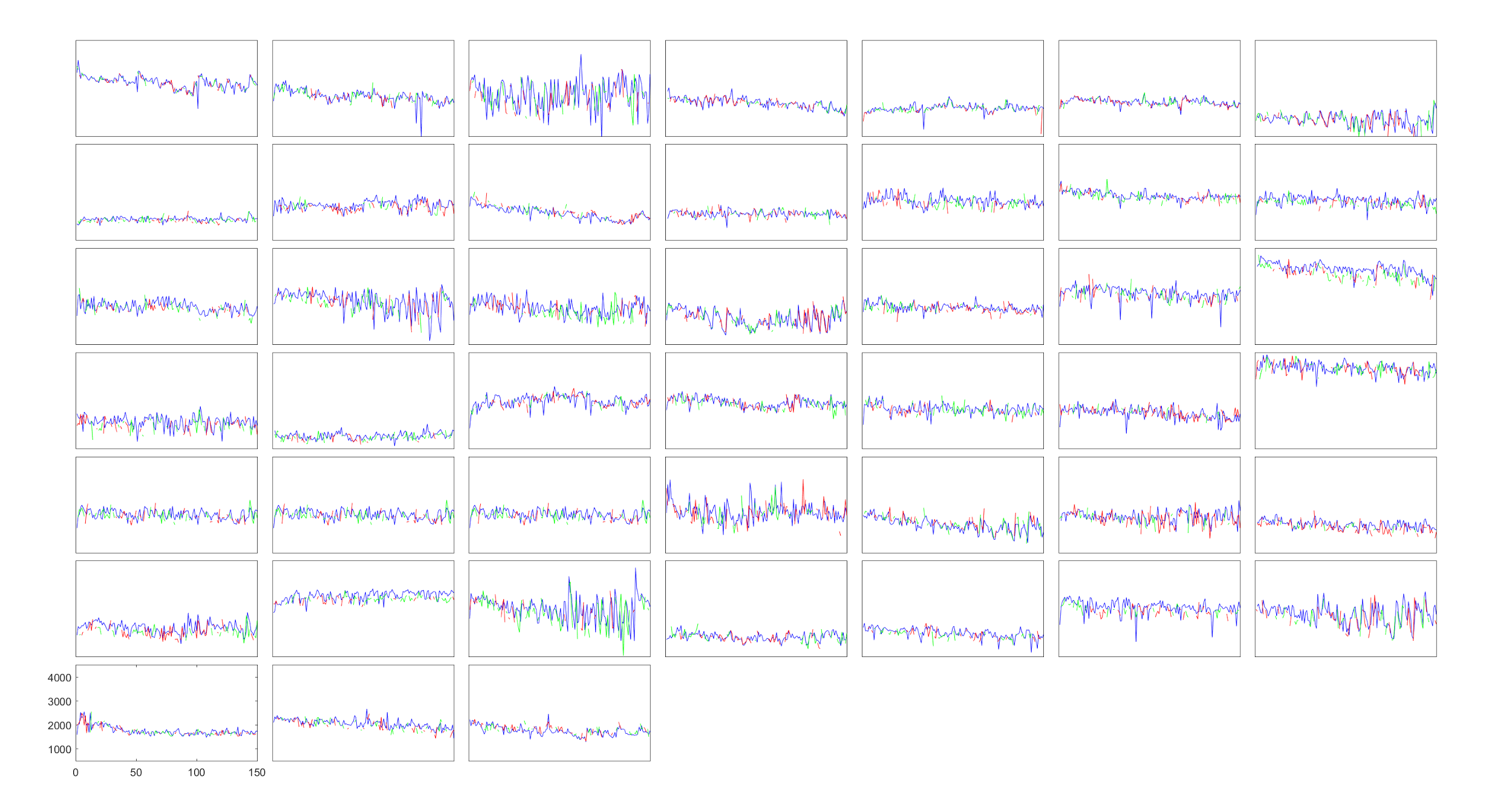
B.

C.

D.

**Figure 2. Blink Removal Process**

*Panel (A) shows a typical smoothed 2500 ms pre-choice pupil trace (anticipatory response removed). (B) Blinks were removed via detection of steep incline/decines of pupil size. (C) Removal of data between beginning and start of a blink, and (D) interpolation to recover the missing data points.*



**Figure 3. Pupil Diameter Across Trials**

## *Each subplot shows a single participant’s average pupil dilation for each trial (x-axis). Blue = first stage, red = second stage; first option, green = second stage; second option. Given the large degree of overlap across stimuli (line colors), the luminance of the stimuli is unlikely to be a large noise variable affecting other aspects of this paper’s analyses. For some participants, a noticeable pupil constriction can be seen near the 51st and 101st trials--potentially due to the fact that working memory has been cleared*

## Behavioral Analysis

Similar to Jepma and Nieuwenhuis (2011)’s approach--in order to classify decisions on a given stage as exploratory or exploitative, the estimated Q-values for each trial and each state were used to predict the optimal decision for the agent. Decisions that followed suit to the value-estimates were considered exploitative, as they represented the agent repeating previously rewarding behaviors. Decisions that aligned with the lower Q-value option were considered exploratory given that such decisions can be thought of as representing a dismissal of one’s current value estimates in favor of seeking a new, longer term, reward.

## Regression Analysis

To understand some of the factors contributing to the variance underlying the dynamics of the pupillary response, a multivariate linear regression analysis was conducted with baseline pupil size preceding a choice as the dependent variable, and the following variables as the model predictors: all hybrid model parameter estimates (*α*, *β*, *λ*, and *w*), cumulative sum of state visits (state cumsum), Q-differences, action outcomes, booleans indicating whether an action following baseline pupil diameter was exploratory or not, and an indicator variable pertaining to the state visited preceding a choice (i.e. the state in which pupil size is measured)

The inclusion of hybrid model parameters allows us to investigate the predictive value of each parameter on pre-choice pupil-size measurements. Given the likelihood that model-based vs model-free behavior might affect the impact of other predictors’ effects on pupil size, the *w* parameter was allowed to interact with all other hybrid model parameters, as well as action outcomes. The inclusion of the ‘state cumsum’ variable allows to control for the observation that pupil size tends to decrease throughout a given trial-set (Fig. 3). Q-differences refer to the absolute difference in a chosen action’s Q-value and the alternative action’s Q-value. Such a variable allows for the conceptualization of the *degree* of exploratory and exploitative decisions. That is, if there is a large absolute Q-difference, then that action is either very exploratory or very exploitative. Action outcomes were included to account for the possibility of an anticipatory pupil response to an upcoming (predicted) reward. For the first action of each trial (stage-1), outcomes of stage-2 were inputted as stage-1 outcomes (rather than zeroes) given the possibility of a reward-prediction response at the beginning of a trial and not necessarily during the second stage action. Booleans indicating whether or not actions were exploratory were included as an interaction term to all (linear) model parameters to account for potential differences in the regression coefficients due to differing behavior types. Finally, an indicator of which state was visited was included due to the concern of differing luminance values of each stimulus/state, as well as to act as an interaction term regarding a possible difference in effect size of trial outcome given the current state (necessary due to the manner in which outcome values from stage-2 are transferred to stage-1 outcome values). The linear model is summarized as follows:

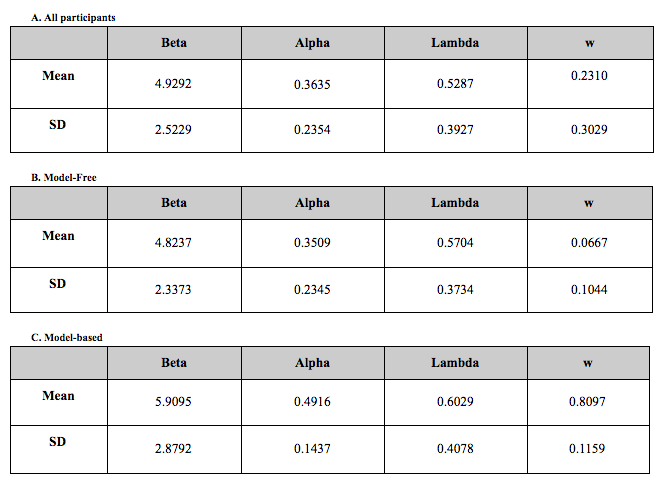
*(8)*

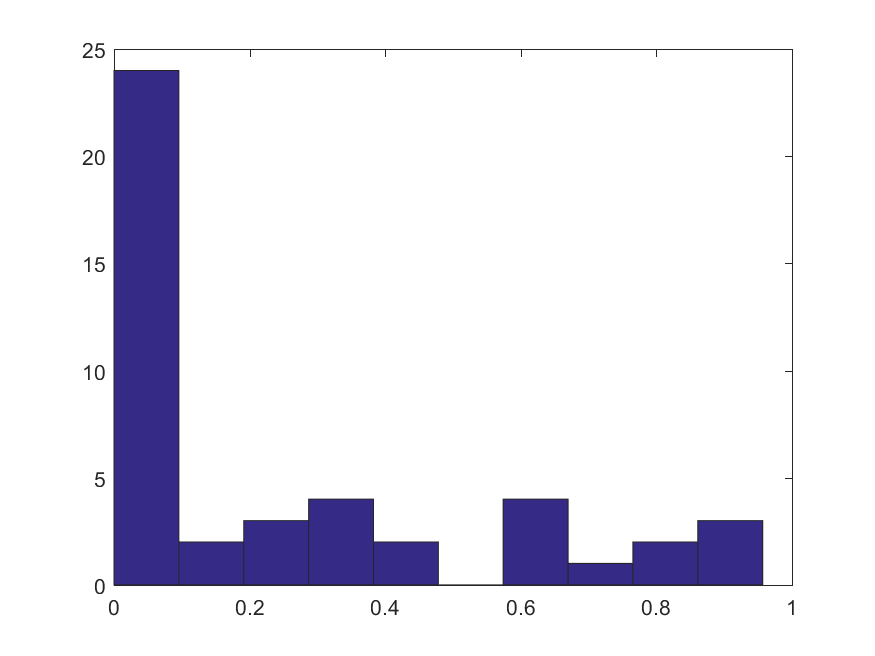
*pre-choicePupilDiams~[(allOutcomes\*allStates\*allWvals)+allCumSums+allQdiffs+allAlphas\*allWvals+allBetas\*allWvals+ …*

*allLambdas\*allWvals]\*allExploreBooleans*

# Results

## Parameter estimates

 After fitting each participant’s data to the core hybrid Q-model, reasonable parameter estimates were recovered (Fig. 4A). Plotting the distribution of *w* values reveals a bias towards model-free learning (lower *w* values, Fig. 4D). In order to convey subgroup parameter estimates, participants were subdivided into *w* value bins corresponding to < .33 (model-free, n=33) and > .66 (model-based, n=7). Average estimates for each subgroup are displayed in Figure 4B-C.



**Figure 4. Average parameter estimates across participants; Distribution of the parameter ‘*w*’ across participants**

*A,B,C: Average parameter estimates for all participants, model-free participants only, and model-based participants only. D. Histogram of w-estimates for all participants. Most participants were found to lean towards a model-free style of learning, as indicated by the high frequency of low value parameter estimates of the ‘w’ parameter. To test differences between learning-styles, participants with a ‘w’ value of > .66 were classified as model-based learners (n=7) and those with a value of < .33 were classified as model-free learners (n=33).*

## Variables underlying pre-choice pupil diameter

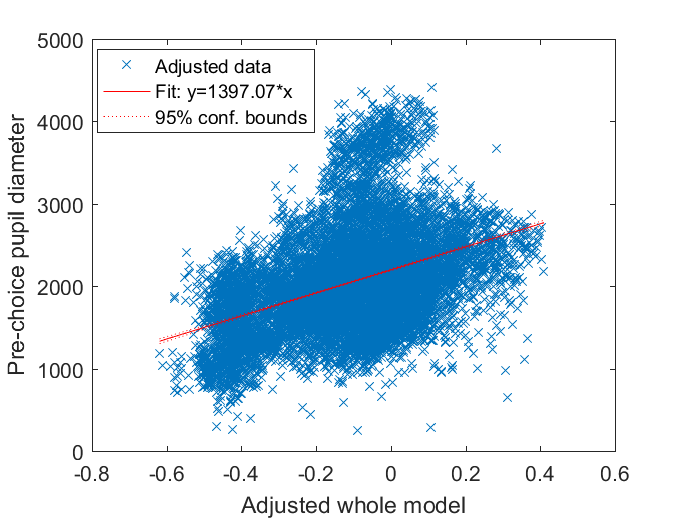
Altogether, the multivariate linear model was able to explain roughly 17% of the variance in the data (Fig. 5). The following effects discussed refer to partial regressors that first hold all other model predictors constant, which biases these results (aside from those pertaining to *w*) towards a summary of model-free learners (given predictors are held constant at their respective means, and model-free learners were predominant in this sample). Beginning with the less exciting predictive variables of the linear regression model, the nominal variable indicating trial state significantly impacted the dilation of the pupil during pre-choice measurements, with blue state decreasing diameter by -84.21 units (scale not given) in comparison to the start state, t(12868) = -3.60, p < .001, and the purple state having a similar estimate of -123.01, t(12868) = -5.11, p < .001. The effect of state exposure across trials also lowered pupil dilation, albeit to a very small degree, by -1.92; t(12868) = -7.19, p < .001.

Moving on to the variance explained by the hybrid model parameters, pre-choice pupil dilation was found to increase with alpha by 712.10 per increment of alpha, t(12868) = 17.45, p < .001; decrease with beta by -45.06 per increment (plus one) of beta, t(12868) = -11.51, p < .001; decrease with lambda by -46.37 per increment of lambda, t(12868) = -2.26, p < .05; and increase by 237.28 per increment of *w,* t(12868) = 3.08, p < .001.

Analysis of behavioral measures revealed a significant difference between exploration and exploitation behavior--exploratory decisions were preceded by an average increase in pupil size (relative to exploitation decisions) of 131.20, t(12868) = 2.31, p < .05. Somewhat surprisingly, Q-differences did not have a significant impact on pupil diameter. Trial outcomes, however, did correlate with pre-choice pupil diameter; with pre-choice diameter decreasing by -51.25 preceding the acquisition of a reward (compared to no reward), t(12868) = -2.72, p < .01. No interaction between the effect of outcome and state was observed, but there was an observed interaction between outcome and learning strategy; model-based learning lead to an increase dilation in response to reward while model-free learners showed pupil constriction in response to a reward (Fig. 8), t(12868) = 2.42, p < .05.

In terms of interactions between exploration booleans and other parameters--increasing alphas were found to increase pupil size in a steeper manner preceding exploitation actions compared to exploration actions (Fig. 6A), t(12868) = -3.24, p < .01, estimate= -245.18; increasing betas were found to decrease pupil size in a steeper manner preceding exploration actions compared to exploitation actions (Fig. 6B), t(12868) = -2.90, p < .01, estimate = -19.63; and increasing *w* values (increasing model-based behaviors) trended towards a decrease and increase in pupil size preceding exploitative and exploratory actions, respectively, compared to model-free learners (Fig. 6C), t(12868) = 1.76, p=.08; estimate=224.98. This particular interaction was also found to be modulated by lambda, with increasing lambda exacerbating the difference between learning strategies, t(12868) = 2.1177, p < .05; estimate=191.98.

In regards to interactions between *w* and other parameters, more model-based learning lead to a decrease in pupil diameter with increasing alphas, and more model-free learning lead to an increase in pupil diameter with increasing alphas (Fig. 7A), t(12868) = -7.32, p < .001; estimate = -735.56. The effect of beta values on pupil size was also modulated by *w*--with more model-based learners showing less of a decline in pupil diameter in response to higher betas than model-free learners (Fig. 7B), t(12868) = 5.92, p < .001, estimate = 50.61. This modulation was reversed in regards to the parameter lambda--model-based learners showed a much stronger decline in pupil size with increasing lambda than model-free learners (Fig. 7C), t(12868) = -15.73, p < .001.

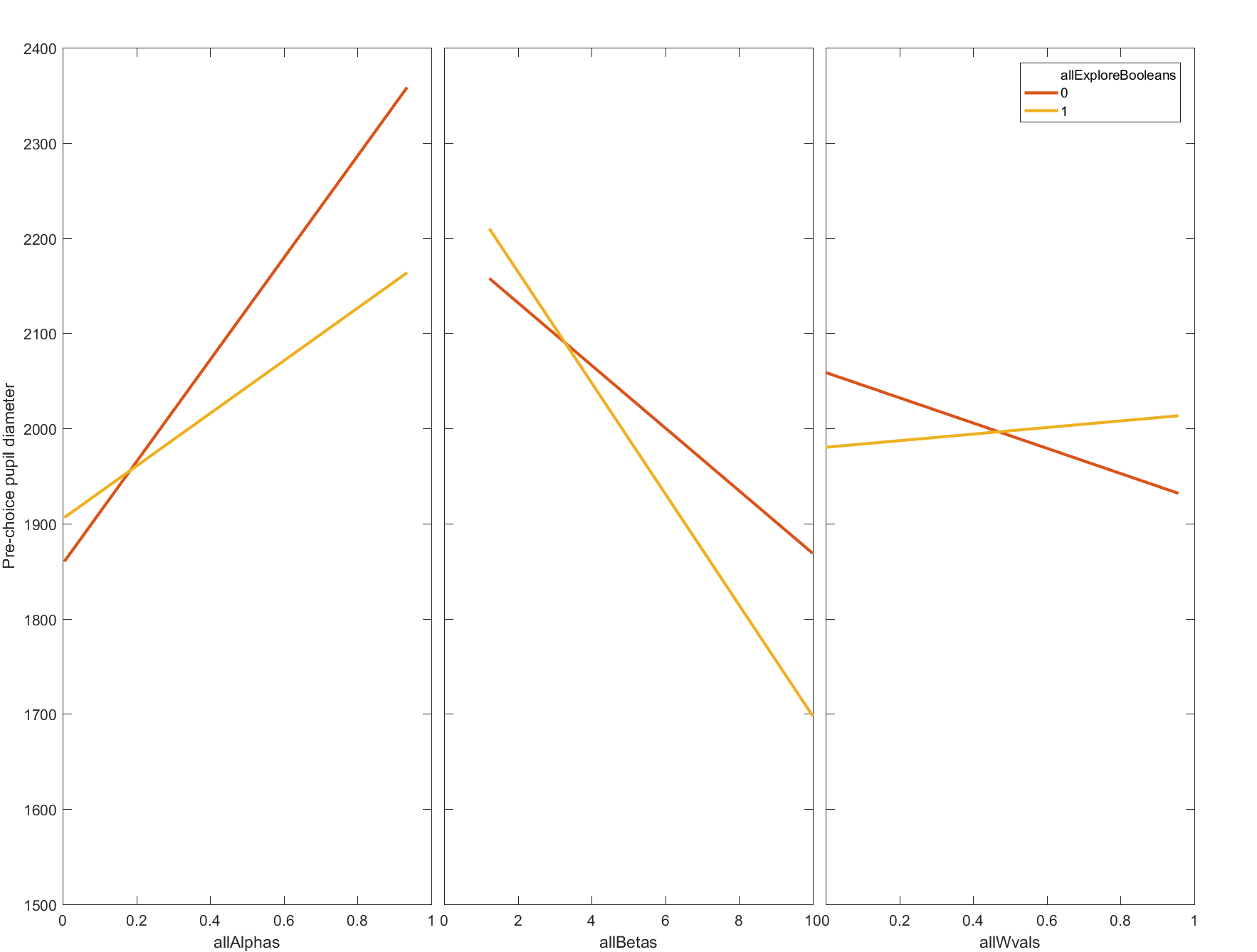


\

**Figure 5. Adjusted whole model predicting single-trial pupil diameters for all participants**

*Illustration predicted variance by multivariate regression model. R-squared: 0.167.*

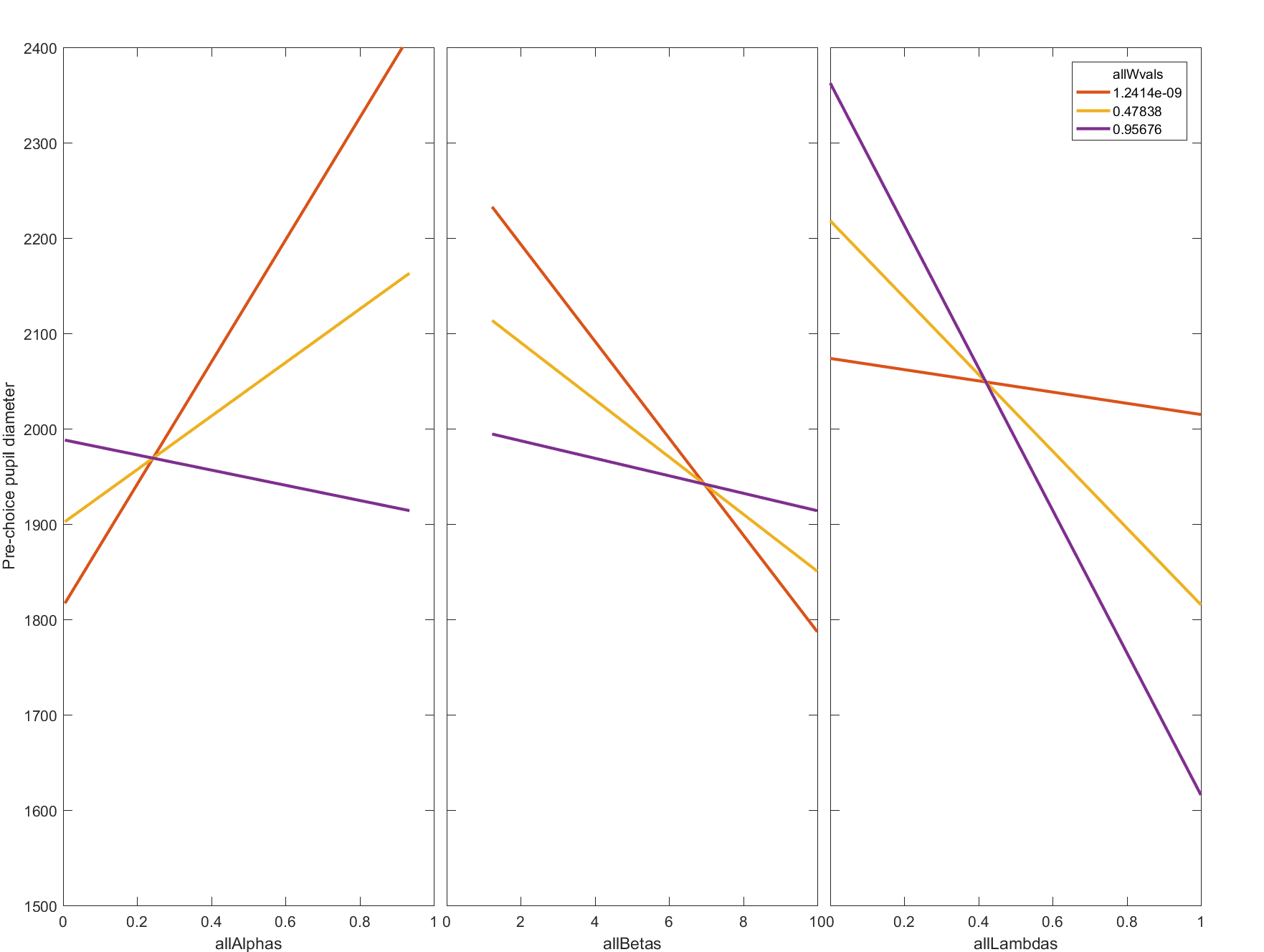
**A. B. C.**



**Figure 6. Interactions between exploration booleans and other model parameters**

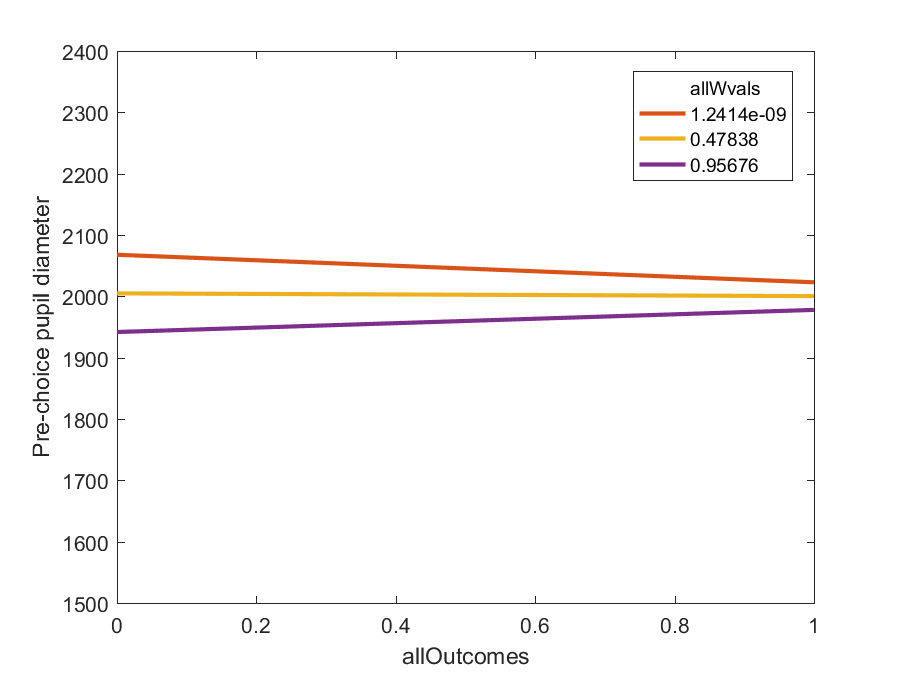
*Plot illustrating interaction effects extracted from the linear model. Exploit = 0 (orange); Explore = 1 (yellow). Only significant or near significant (allWvals) interactions with booleans are shown. The overall slopes of panels A&B must be interpreted with caution given that they implicitly largely reflect only differences within model-free learning strategies given that the majority of participants used a model-free learning strategy.*

**A. B. C.**



**Figure 7. Interactions between *w* (model-based vs model-free) and other model parameters**

*Plot illustrating interactions between w and: alpha, beta, and lambda. A value of w nearer to 1 indicates more model-based learning (purple), and a value nearer to 0 indicates more model-free learning (orange).*



**Figure 8. Interaction between learning strategy and reward response**

*Model-based learners show increased dilation in response to a reward, while model-free learners show the opposite relationship.*

# Discussion and Future Work

## Analysis predictions revisited

All other predictors held constant (which biases this summary to pertain to model-free learners), increasing alpha predicts a larger pupil diameter, and this effect is more dramatically observed for exploitative decisions than exploratory decisions. If one interprets alpha as a variable signifying the extent to which the agent incorporates novel information to adjust its value-estimates of the world, then increasing alpha could be expected to increase pupil diameter in the sense that the pupil might act as a gatekeeper of sorts for allowing in novel information to be used for sensory-model updating. Such a conceptualization predicts an increased pupil diameter throughout exploratory trials, which is also observed from the regression analysis. Regarding the interaction between alpha and exploration booleans, the proposal that exploitative decisions are likely to cue memory retrieval processes and typically lower pupil size, as is observed in this model, implies that the observed interaction can be understood to stem partially from the difference, at low alpha values, between typical exploratory or exploitative processes cueing learning or memory processes, respectively. That is, when alpha increases, it’s feasible to expect the agent to engage in sensory model-updating even in familiar states (exploitative decisions), which would dramatically increase the average pupil diameter of exploitative decisions compared to an agent that solely engages in recollection during exploitative decisions. Following this same logic, a nearly completely opposite effect is observed regarding the parameter beta. That is, as beta increases (and exploitation increases with it), pupil diameter tends to decrease, and this decrease is more dramatically observed for exploratory decisions than exploitative decisions. Again, this could be due to an increased propensity of recollection processing taking place during exploratory decisions (e.g. the agent is stuck in the past and processing memory storage even when it encounters a novel environment).

Of particular interest to this paper was the impact of the variable *w* on pre-choice pupil responses preceding both exploratory and exploitative decisions. While the interaction between *w* and explore booleans did not reach significance, it trended towards the predicted relationship (p = .08), with model-free learners exhibiting larger pupil dilations than model-based learners in terms of exploitative decisions, and smaller pupil dilations than model-based learners in terms of exploratory decisions (Fig. 6C). Regarding the exploitative condition, this evidence falls in line with the idea that model-based learners engage more in sensory-state recollections and therefore exhibit smaller pupil diameters preceding exploitative decisions than model-free learners due to a more probable influx of sharp wave ripples, which have been found to occur more frequently during familiar trials (Leonard & Hoffman, 2017). In further support of this idea, another relevant piece of evidence lies in the significant interaction between lambda and *w*. As can be seen in Figure 7C, as lambda increases, pupil size drastically decreases for model-based participants compared to model-free learners. Increasing lmabda also exacerbates the interaction between *w* and explore booleans (see Results). Given that lambda controls for how relevant the second stage prediction error is when updating stage-1’s action-value, it seems entirely plausible that this evidence collectively corresponds to recollection of stage-2 state-transition predictions when updating stage-1’s Q-value, which manifests as a larger decrease in pupil diameter than in model-free participants during exploitation trials. Use of this information during Q-updating could also explain why increasing alphas corresponds with decreasing pupil diameter for model based-learners (Fig. 7A).

All in all, these findings suggest that increased memory processing appears to be evident amongst model-based learners. It should of course be mentioned, however, that model-free learners do, indeed, exhibit smaller pupil sizes preceding exploitative decisions compared to exploratory decisions--just as model-based learners exhibit--suggesting that both learning strategies require for memory processing of some sort. In fact, model-free learners even appear to show a more dramatic decline in pupil size when engaged with very exploitative behavior (i.e. high beta values) compared to model-based learners (Fig. 7B, though take note of the large differences at lower beta values as well). The present results, therefore, suggest that pupil constriction is either more frequent or more dramatic in a model-based framework. Regarding the exploratory condition--the current findings also elucidate either a larger or more frequent dilation of the pupils when model-based learners encounter novel information, compared to model-free learners. This effect could reasonably reflect increased information uptake by model-based learners, or similarly, a larger working-memory load, which has been found to correlate with increased pupil diameters, due to storage of state-transition details (Peysakhovich et al., 2015; Tsukahara et al., 2016).

## Final remarks and future steps

This study was a successful exploratory examination of the pupillary response in relation to reinforcement learning parameters that allows for a fairly intuitive conceptualization of many findings pertaining to the pupillary response throughout decision making processes. One particular weakness that became evident during analysis, however, was the lack of sufficiently dynamic parameters to capture the each participants’ behavior. Redoing the analysis, I would treat each chunk of 50 trials for each participant as a dataset to fit to a unique set of parameters--allowing for a higher degree of precision when fitting the data to a linear model. Aside from this improvement, however, there are a great number of ways to continue this direction of research. For starters, it would be good to verify that human SWRs correlate strongly with pupil constriction as observed in mice. Such a verification would help to strengthen confidence in the present paper’s interpretations of available evidence. Furthermore, it will be necessary to further refine the exact relationship between pupil diameter and memory processes. The present paper remained intentionally vague regarding which kinds of memory processes take place during pupil constriction due to the lack of current progress in classifying the exact functionality of sharp wave ripples in different contexts. It’s quite possible, however, that certain degrees/rates of pupil constrictions might be able to distinguish between replay, preplay, and other memory-related processes. A model-learning approach could feasibly help to clarify such mechanisms. Such an analysis has already been initiated by the present author, with certain memory processes controlled by pupillometry measurements on a trial-to-trial basis showing better model fits (BIC values) than others. Reassuring to the current conclusions in this paper, most pupillometry-controlled models tested thus far show improved BIC values compared to fits that neglect to exploit such additional single-trial resolution information--demonstrating that the proposed interpretation of pupillary dynamics is at least useful for predictive purposes. In addition, the current paper’s conceptualization of the pupillary response may help to further elucidate the exact nature of model-based versus model-free learning strategies, including whether or not such strategies truly operate in a mutually exclusive manner. The distinct pupillometry correlates of both learning strategies presented here further support the notion that such strategies are distinct, but there is still room for much more debate on this issue. One additional jumping off point may be to further investigate the reason why model-based learners showed a different response to rewards than model-free learners did (Fig. 8). Finally, if this paper’s findings are further validated by future research, they suggest that a simple eye-tracker could be used to extract measurements of typically closed-off internal processes pertaining to sensory and memory processing. Such a measurement may be extremely favorable for many applications which, for one reason or another, cannot practically use an EEG for measurements of such processes.

# Acknowledgements

The analyses in this paper would not be possible were it not for the generosity and willingness to share data from Konovalov and Krabjich. Furthermore, these authors were more than willing to share with me several tips regarding the model-fitting procedure in this field of literature. Their assistance is greatly appreciated.

# References

Abou Elseoud, A., Nissila, J., Liettu, A., Remes, J., Jokelainen, J., Takala, T., . . . Kiviniemi, V. (2014). Altered Resting-State Activity in Seasonal Affective Disorder. *Human Brain Mapping, 35*(1), 161-172. doi:10.1002/hbm.22164

Atherton, L. A., Dupret, D., & Mellor, J. R. (2015). Memory trace replay: the shaping of memory consolidation by neuromodulation. *Trends in neurosciences, 38*(9), 560. doi:10.1016/j.tins.2015.07.004

Bitsios, P., Prettyman, R., & Szabadi, E. (1996). Changes in autonomic function with age: a study of pupillary kinetics in healthy young and old people. *Age and ageing, 25*(6), 432. doi:10.1093/ageing/25.6.432

Buhry, L., Azizi, A., & Cheng, S. (2011). Reactivation, Replay, and Preplay: How It Might All Fit Together. In *Neural. Plast.* (Vol. 2011).

Buzsáki, G. (2015). Hippocampal sharp wave‐ ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus, 25*(10), 1073-1188. doi:10.1002/hipo.22488

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience, 35*(7), 1024-1035. doi:10.1111/j.1460-9568.2011.07980.x

David, J. F., & Matthew, A. W. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature, 440*(7084), 680. doi:10.1038/nature04587

Daw, N. D. (2011). *Trial-by- trial data analysis using computational models*.

Daw, Nathaniel d., Gershman, Samuel j., Seymour, B., Dayan, P., & Dolan, Raymond j. (2011). Model- Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron, 69*(6), 1204-1215. doi:10.1016/j.neuron.2011.02.027

Ganis, G., Thompson, W. L., & Kosslyn, S. M. (2004). Brain areas underlying visual mental imagery and visual perception: an fMRI study. *Cognitive Brain Research, 20*(2), 226-241. doi:10.1016/j.cogbrainres.2004.02.012

Gläscher, J., Daw, N., Dayan, P., amp, Amp, Apos, & Doherty, J. P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model- Based and Model- Free Reinforcement Learning. *Neuron, 66*(4), 585-595. doi:10.1016/j.neuron.2010.04.016

Houillon, A., Lorenz, R. C., Boehmer, W., Rapp, M. A., Heinz, A., Gallinat, J., & Obermayer, K. (2013). The effect of novelty on reinforcement learning. In *Prog. Brain Res.* (Vol. 202, pp. 415-439).

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

Kang, M.-S., Hong, S., Blake, R., & Woodman, G. (2011). Visual working memory contaminates perception. *Psychonomic Bulletin & Review, 18*(5), 860-869. doi:10.3758/s13423-011-0126-5

Karl, F. (2010). The free- energy principle: a unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127. doi:10.1038/nrn2787

Koenig, S., Uengoer, M., & Lachnit, H. (2017). Pupil dilation indicates the coding of past prediction errors: Evidence for attentional learning theory. *Psychophysiology*. doi:10.1111/psyp.13020

Lavín, C., Martín, R. S., & Jubal, E. R. (2014). Pupil dilation signals uncertainty and surprise in a learning gambling task. *Frontiers in Behavioral Neuroscience, 7*. doi:10.3389/fnbeh.2013.00218

Lee, D., Seo, H., & Jung, M. W. (2012). Neural Basis of Reinforcement Learning and Decision Making. *Annu. Rev. Neurosci., 35*, 287-308. doi:10.1146/annurev-neuro-062111-150512

Leonard, T. K., & Hoffman, K. L. (2017). Sharp- Wave Ripples in Primates Are Enhanced near Remembered Visual Objects. *Current Biology, 27*(2), 257-262. doi:10.1016/j.cub.2016.11.027

McGinley, Matthew j., David, Stephen v., & McCormick, David a. (2015). Cortical Membrane Potential Signature of Optimal States for Sensory Signal Detection. *Neuron, 87*(1), 179-192. doi:10.1016/j.neuron.2015.05.038

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning: dissecting multiple reinforcement- learning systems by taxing the central executive. *Psychological science, 24*(5), 751. doi:10.1177/0956797612463080

Pascale, W., Anthony, D., & Wolfram, S. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature, 412*(6842), 43. doi:10.1038/35083500

Peysakhovich, V., Dehais, F., & Causse, M. (2015). Pupil Diameter as a Measure of Cognitive Load during Auditory- visual Interference in a Simple Piloting Task. *Procedia Manufacturing, 3*, 5199-5205. doi:10.1016/j.promfg.2015.07.583

Preuschoff, K., T Hart, B. M., & Einhäuser, W. (2011). Pupil Dilation Signals Surprise: Evidence for Noradrenaline's Role in Decision Making. *Frontiers in neuroscience, 5*, 115. doi:10.3389/fnins.2011.00115

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition, 112*, 92-97. doi:10.1016/j.bandc.2015.11.003

Tsukahara, J. S., Harrison, T. L., & Engle, R. W. (2016). The relationship between baseline pupil size and intelligence. *Cognitive Psychology, 91*, 109-123. doi:10.1016/j.cogpsych.2016.10.001

Volodymyr, M., Koray, K., David, S., Andrei, A. R., Joel, V., Marc, G. B., . . . Demis, H. (2015). Human- level control through deep reinforcement learning. *Nature, 518*(7540), 529. doi:10.1038/nature14236

Wiegand, J.-P. L., Gray, D. T., Schimanski, L. A., Lipa, P., Barnes, C. A., & Cowen, S. L. (2016). Age Is Associated with Reduced Sharp- Wave Ripple Frequency and Altered Patterns of Neuronal Variability. *The Journal of neuroscience : the official journal of the Society for Neuroscience, 36*(20), 5650. doi:10.1523/JNEUROSCI.3069-15.2016