# Nonparametric Learning Rules from Bandit Experiments: The Eyes have it![*]

Yingyao Hu Johns Hopkins

Yutaka Kayaba Caltech

Matt Shum Caltech

First draft: April 2010
This draft: June 7, 2010

**Abstract**

We estimate nonparametric learning rules using data from dynamic two-armed bandit (probabilistic reversal learning) experiments, supplemented with auxiliary measures of subjects' beliefs, in the form of their eye-movements during the learning experiment. We apply recent econometric developments in the estimation of dynamic models. The estimated choice probabilities and learning rules from our nonparametric models have some distinctive features; notably that subjects tend to update in a non-smooth manner following positive "exploitative" choices (those made in accordance with current beliefs). Simulation results show that the beliefs implied by the nonparametric learning rules are more similar to those from a reinforcement learning model, than a Bayesian learning model.

## 1 Introduction

How do individuals learn from past experience in dynamic choice environments? We address this question by presenting nonparametric estimates of subjects' learning rules in a dynamic two-armed bandit (probabilistic reversal learning) problem where subjects must repeatedly

[*]We are indebted to Antonio Rangel for his encouragement and for the funding and use of facilities in his lab. We thank Dan Ackerberg, Peter Bossaerts, Colin Camerer, Cary Frydman, Ian Krajbich, Pietro Ortoleva, and participants in presentations at U. Arizona, Caltech, UCLA, U. Washington and Choice Symposium 2010 (Key Largo) for comments and suggestions.

guess which of the two arms yields a (stochastically) higher reward. Auxiliary measures of subjects' eye movements as they make their choices are employed to "pin down" subjects' beliefs in each round of the learning experiment. To our knowledge, the nonparametric estimation of learning models is a new endeavor in both the behavioral learning literature, as well as the empirical literature in economics and marketing in which dynamic learning models are estimated structurally. Estimating the learning rules nonparametrically allows us to compare competing learning models in a manner quite distinctive than that taken in the existing literature.

Recently, a number of researchers in industrial organization and marketing have estimated learning-based models of dynamic choice. Some representative papers include Ackerberg (2003), Erdem and Keane (1996), Crawford and Shum (2005), Chan and Hamilton (2006) and Marcoul and Weninger (2008). This literature typically assumes that agents process information according to a forward-looking Bayesian learning model. This restrictive assumption is driven in part by data considerations: oftentimes, all that is observed are the sequences of agents' choices, so that a lot of (parametric) structure must be placed on the learning model for identification.

In controlled experimental settings, richer data are observed: not only subjects' choices, but also the outcomes (rewards) from their choices. In addition, depending on the laboratory setting, there is also the opportunity to observe "auxiliary" measures of subjects' beliefs (or valuations), such as brain activity (cf. Yoshida and Ishii (2006), Boorman, Behrens, Woolrich, and Rushworth (2009) in the recent fMRI neuroscience literature) or eye movements (as in the present paper)[1]. Because of this data richness, researchers in the behavioral/experimental literature have been able to consider more flexible learning rules, and to test the fully-rational Bayesian learning benchmark versus boundedly-rational, backward-looking "reinforcement learning" (RL) rules (cf. Sutton and Barto (1998)). An incomplete list of papers includes Grether (1992), El-Gamal and Grether (1995), Charness and Levin (2005), Kuhnen and Knutson (2008), and Payzan and Bossaerts (2009). Particularly, RL has attracted considerable attention in the recent neuroeconomics and decision neuroscience literature (cf. Glimcher, Camerer, Poldrack, and Fehr (2008), Rushworth and Behrens (2008)), ever since studies showing that the "prediction errors" of these models are apparently encoded in certain areas of the brain (cf. Schultz, Dayan, and Montague (1997)) for evidence from primates). Recently, RL models have also been used to explain

[1] Although it is not about beliefs, Wang, Spezio, and Camerer (forthcoming) use eye-tracking data to identify unobservable variables (truth-telling and deception) in sender-receiver games.

some observed anomalies in savings and investment behavior (eg. Choi, Laibson, Madrian, and Metrick (2009), Odean, Strahilevitz, and Barber (2004)).[2]

In this paper, we take a new approach to assessing learning in experimental settings. Taking advantage of recent developments in the econometrics of estimating dynamic models with serially-correlated unobservables, we use the observed experimental and auxiliary data to estimate, nonparametrically, subjects' choice probabilities and learning rules, without imposing *a priori* functional forms on these functions. Thus, our learning rules can be reasonably interpreted as "what the subjects actually think", as reflected in their observed choices. Subsequently, we compare our estimated learning rules to specific parameterized learning rules which have been considered in the previous literature, including the Bayesian and reinforcement-learning models.

Moreover, we estimate not only the learning rules nonparametrically, but also the choice probabilities. Choice probabilities are key parameters in machine learning and decision neuroscience models (cf. Sutton and Barto (1998), Daw, O'Doherty, Dayan, Seymour, and Dolan (2006), Doya (2002)). Although parameterized models of choice behavior have been examined in several studies (cf. Daw, O'Doherty, Dayan, Seymour, and Dolan (2006)), to our knowledge, this research would be the first to examine choice behavior without imposing *a priori* functional forms on the choice probabilities.

Our approach differs from a common *modus operandi* in the behavioral/experimental literature, which has been to use the observed choice data from the experiment to calibrate parameters for competing learning models. Subsequently, the competing learning models are simulated, and verification is based upon comparing the simulated learning rules with the observed auxiliary belief measurements. For instance, Hampton, Bossaerts, and O'Doherty (2006) test between a Bayesian and reinforcement-learning model on the basis of two-armed bandit experiments supplemented with brain activity information from fMRI brainscans. Other papers utilizing a similar methodological framework include Behrens, Woolrich, Walton, and Rushworth (2007), Boorman, Behrens, Woolrich, and Rushworth (2009), Daw, O'Doherty, Dayan, Seymour, and Dolan (2006), Yoshida and Ishii (2006).

Methodologically, this paper represents a novel application of econometric tools recently developed for the estimation of nonclassical measurement error models and dynamic discrete-choice models (Hu (2008), Hu and Shum (2008)). Because subjects' underlying beliefs are

[2] In the computational IO literature, such learning algorithms have also been used to ease the computational burden associated with dynamic equilibrium models, cf. Pakes and McGuire (2001), Imai, Jain, and Ching (2009).

unobserved and also serially correlated over time, the learning model is a particular case of a nonlinear "hidden Markov" model, which are challenging to estimate (Ghahramani (2001)). Our approach is to fit the learning model into a dynamic misclassification framework, in which the eye-movement measures play the role of "noisy measurements" of the underlying belief process.[3]

In Section 2, we describe the dynamic two-armed bandit learning (probabilistic reversal learning) experiment, and the eye movement data gathered by the eye-tracker machine. In Section 3, we present an econometric model of subjects' choices in the bandit model, and discuss nonparametric identification. We also describe our estimation procedure there. In Section 4, we describe the experimental data, and present our nonparametric estimates of subjects decision rules and learning rules. Section 5 contains a comparison of our estimated learning rules to "standard" learning rules, including those from the Bayesian and reinforcement-learning models. Section 6 concludes.

## 2 Two-armed bandit learning (probabilistic reversal learning) experiment

The learning experiments considered in this paper follow the setup in Hampton, Bossaerts, and O'Doherty (2006). We consider an experiment where subjects chooses between two actions, called "blue" and "green", where the rewards of these two actions are changing across trials.

In each trial $t$, a subject chooses one of two actions (which we call interchangeably "arms" or "slot machines" in what follows): $Y_t \in \{1(=\text{"green"}), 2(=\text{"blue"})\}$. Which of these arms is the "good" one varies trial-by-trial, as described by the state variable $S_t \in \{1, 2\}$. The state variable is never observed by subjects. When $S_t = 1$, then green (blue) is the "good" ("bad") state, whereas if $S_t = 2$, then blue (green) is the "good" ("bad") state.

The rewards $R_t$ that the subject receives in trial $t$ depends on the action taken, as well as (stochastically) on the current state: the good (bad) arm yields rewards

$$R_t = \begin{cases} \text{“}2''(= \$0.50) & \text{with prob } 0.7 \ (0.4) \\ \text{“}1''(= -\$0.50) & \text{with prob } 0.3 \ (0.6) \end{cases} \tag{1}$$

[3] Relatedly, Samejima, Doya, Ueda, and Kimura (2004) consider Bayesian estimation of a reinforcement learning model using sequential Monte Carlo ("particle filtering") methods.

The state evolves according to an exogenous binary Markov process. At the beginning of each block, the initial state $S_1 \in \{1, 2\}$ is chosen with probability 0.5, randomly across all subjects and all blocks. Subsequently, the state evolves with transition probabilities

| $P(S_{t+1}\|S_t)$ | $S_t = 1$ | $S_t = 2$ |
|---|---|---|
| $S_{t+1} = 1$ | 0.85 | 0.15 |
| $S_{t+1} = 2$ | 0.15 | 0.85 |

Because $S_t$ is not observed by subjects, and is serially-correlated over time, there is the opportunity for subjects to learn and update their beliefs about the current state on the basis of past rewards. The goal of the exercise in this paper is to infer subjects' learning (that is, belief updating) rule, on the basis of their observed choices.

**Remark 1 (reversal learning):** *This bandit problem with reversal learning differs in important ways from the "standard" multi-armed bandit problem (cf. Gittins and Jones (1974), Banks and Sundarum (1992)), in which the states of the bandits are fixed over all periods and the bandits are "independent" in that a reward from one bandit is uninformative about the state of another bandit. The optimal Bayesian decision rule features exploration (or "experimentation"), which recommends sacrificing current rewards to achieve longer-term payoffs.*[4] *In the setting considered in this paper, however, the bandits are negatively correlated, so that positive information about one slot machine implies negative information about the other. This should largely diminish the incentive for subjects to experiment.*

## 2.1 Data

The experiments were run over several weeks time in November-December 2009. We used 21 subjects, recruited from the Caltech Social Science Experimental Laboratory (SSEL) subject pool consisting of undergraduate/graduate students, post-doctoral students, and community members,[5] each playing for 200 rounds (broken up into 8 blocks of 25 trials). Most of the subjects finished the whole task within 40 minutes, including instruction and practice sessions. Subjects were paid a fixed show-up fee ($20), in addition to the amount earned during the experiment, which was $14.2 in average.[6]

[4] See Crawford and Shum (2005) for an empirical analysis of this phenomenon in pharmaceutical drug demand.

[5] Community members consisted of spouses of students at either Caltech or Pasadena City College (a nearby community college).

[6] For comparison, purely random choices would have earned $10 on average.

Subjects were informed of the reward structure for good and bad slot machines, and the Markov transition probabilities for state transitions (reversals), but were not informed which state was occuring in each trial. For each subject, and each round $t$, we observe the data $(Y_t, S_t, R_t)$. In Figure 1, we present the time line and some screenshots from the experiment. In addition, while performing the experiment, the subjects were attached to an eye-tracker machine, which recorded their eye movements. From this, we constructed the auxiliary variable $Z_t$, which measures the fraction of the reaction time (the time between the onset of a new round after fixation, and the subject's choice in that round) spent gazing at the picture of the "blue" slot machine on the computer screen.[7]

## 3 Econometric model

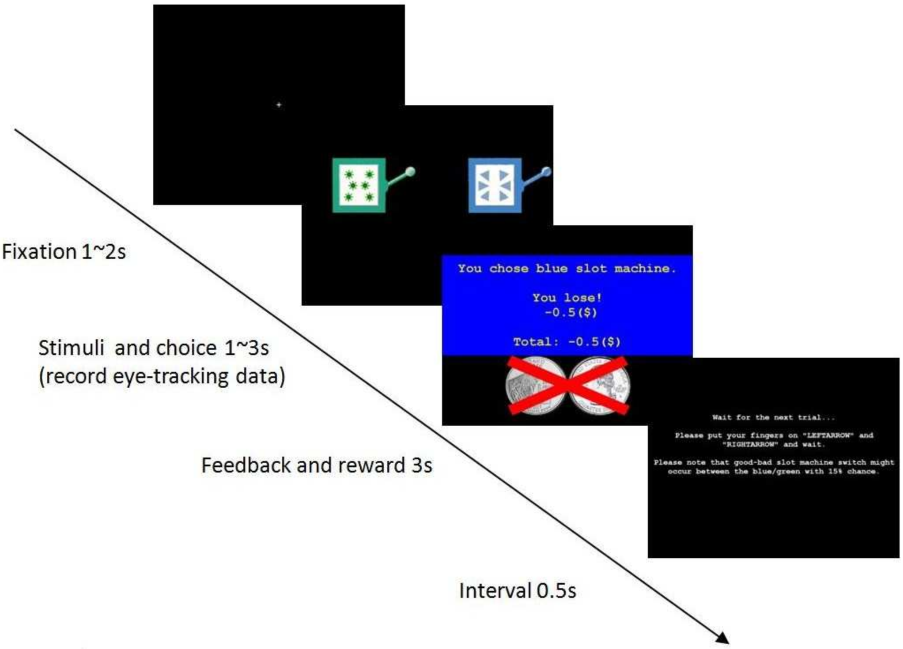
In this section, we describe our econometric model of dynamic decision-making in the two-armed bandit (probabilistic reversal learning) experiment described above, and also discuss the identification and estimation of this model. We introduce the variable $X_t^*$, which denotes the agent's round $t$ beliefs about the current state $S_t$; obviously, agents know their beliefs $X_t^*$, but these are unobserved by the researcher. In what follows, we assume that both $X^*$ and $Z$ are discrete, and take support on $K$ distinct values which, without loss of generality, we denote $\{1, 2, \ldots, K\}$. We make the following assumptions regarding the subjects' learning and decision rules:

**Assumption 1** *Subjects' choice probabilities $P(Y_t|X_t^*)$ only depend on current beliefs. Moreover, the choice probabilities $P(Y_t = y|X^*)$ varies across different values of $X_t^*$ (ie. beliefs affect actions), for $y \in \{1, 2\}$.*

**Assumption 2** *The law of motion for $X_t^*$, which describes how subjects' beliefs change over time given the past actions and rewards, is called the* ***learning rule****. This is a controlled first-order Markov process, with transition probabilities $P(X_t^*|X_{t-1}^*, R_{t-1}, Y_{t-1})$.*

These two assumptions pose very little loss in generality, and hold for most varieties of Bayesian as well as reinforcement learning models.

[7] Across trials, the location of the "blue" and "green" slot machines were randomized, so that the same color is not always located on the same side of the computer screen. This controls for any "right side bias" which may be present (see discussion further below).



You chose blue slot machine.

You lose!
-0.5($)

Total: -0.5($)

Wait for the next trial...

Please put your fingers on "LEFTARROW" and "RIGHTARROW" and wait.

Please note that good-bad slot machine switch might occur between the blue/green with 15% chance.

Figure 1: Timeline of a trial

After a fixation on the cross (top screen), two slot machines are presented (the left-right position is randomized; second screen). Subjects' eye-movements are recorded by the eye-tracking machine here. Subjects choose by pressing the left (right) arrow key to indicate a choice of the left (right) slot machine. After choosing (third screen), a positive reward (depicted by two quarters) or negative reward (two quarters covered by a red X) is delivered, along with feedback about the subject's choice highlighted against a background color corresponding to the choice. In the bottom screen, a subject is transitioned to the next trial, and reminded that the a slot machine may switch from "good" to "bad" (and vice versa) with probability 15%.

**Assumption 3** *The auxiliary measure $Z_t$ is a noisy measure of beliefs $X_t^*$, with the measurement probabilities $P(Z_t|X_t^*)$. We assume that:*
*(i) For all $t$, the $K \times K$ matrix $\mathbf{G}_{Z_t|Z_{t-1}}$, with the $(i,j)$-th entry equal to $P(Z_t = i|Z_{t-1} = j)$, is invertible.*
*(ii) $E[Z_t|X_t^*]$ is increasing in $X_t^*$.*

The invertibility assumption 3(i) is made on the observed matrix $\mathbf{G}_{Z_t|Z_{t-1}}$ with elements equal to the conditional distribution of $Z_t|Z_{t-1}$. Assumption 3(ii) "normalizes" the beliefs $X_t^*$ in the sense that, because large values of $Z_t$ imply that the subject gazed longer at blue, the monotonicity assumption implies that larger values of $X_t^*$ denote more "positive" beliefs that the current state is blue.[8]

The final assumption justifies pooling the data across all subjects and trials for estimating the model:

**Assumption 4** *The choice probabilities $P(Y_t|X_t^*)$, learning rules $P(X_t^*|X_{t-1}^*, R_{t-1}, Y_{t-1})$, and measurement probabilities $P(Z_t|X_t^*)$ are the same for all subjects, trials, and blocks $t$.*

**Remark 2 (stationary in learning models):** *An important benefit of considering a "probabilistic reversal" model (in which the identity of the "good" slot machine changes stochastically across trials) rather than the simpler standard multi-armed bandit model (in which the identity of the "good" arm is fixed across all trials) is that in the latter case, the subject's uncertainty regarding the identity of the "good" arm is decreasing across trials, so that learning rule must also condition on some measure of the subject's uncertainty (such as the number of times a particular arm has been pulled before a given trial) in order to satisfy the stationarity Assumption 4.*[9] *In a probabilistic reversal setting, however, a subject's uncertainty does not decrease across trials. This is an attractive feature because, in our nonparametric estimation approach, conditioning on additional variables decreases the precision of the estimates.*

Given these assumptions, we next describe the nonparametric identification argument.

[8] The model can be easily extended to allow for conditional serial correlation in the auxiliary measure $Z_t$, ie. allowing for a law of motion $P(Z_t|X_t^*, Z_{t-1})$. For $Z_t$ as a measure of eye-movements, as in this paper, the conditional independence assumption across trials appears reasonable, especially given the imposed fixation at the beginning and end of each trial (cf. Figure 1). However, for auxiliary measures in other settings (such as brain activity for fMRI studies), conditional dependence seems more realistic.

[9] For empirical applications of such learning rules in the Bayesian setting, see Ackerberg (2003) or Crawford and Shum (2005).

### 3.1 Nonparametric identification

In this section, we will use the shorthand notation $f(\cdots)$ to denote generically a probability distribution. For identification, we exploit the following relationship: conditional on $(R_{t-1})$, we have

$$f(Y_t, Z_t, X_t^* | Y_{<t}, Z_{<t}, R_{<t}, X_{<t}^*) = f(Y_t, Z_t, X_t^* | Y_{t-1}, R_{t-1}, X_{t-1}^*). \tag{2}$$

Abusing terminology somewhat, we call this a "first-order Markov" property. This is because:

$$\begin{aligned}
&f(Y_t, Z_t, X_t^* | Y_{<t}, Z_{<t}, R_{<t}, X_{<t}^*)\\
=&f(Y_t | Z_t, X_t^*, Y_{<t}, Z_{<t}, R_{<t}, X_{<t}^*) \cdot f(Z_t | X_t^*, Y_{<t}, Z_{<t}, R_{<t}, X_{<t}^*) \cdot f(X_t^* | Y_{<t}, Z_{<t}, R_{<t}, X_{<t}^*)\\
=&f(Y_t | X_t^*) \cdot f(Z_t | X_t^*) \cdot f(X_t^* | X_{t-1}^*, R_{t-1}, Y_{t-1})\\
=&f(Y_t, Z_t, X_t^* | Y_{t-1}, R_{t-1}, X_{t-1}^*).
\end{aligned} \tag{3}$$

In the above, the second equality applies Assumptions 1, 2, and 3.

Consider the joint density $f(Z_t, Y_t | Z_{t-1})$, which is solely a function of variables observed in the data. The unknown functions we want to identify and estimate are:
(i) $f(Y_t | X_t^*)$, the conditional choice probability;
(ii) the learning rule $f(X_t^* | X_{t-1}^*, Y_{t-1}, R_{t-1})$; and
(iii) $f(Z_t | X_t^*)$, the mapping between the auxiliary measure $Z_t$ and the unobserved state $X_t^*$.

The nonparametric identification of these elements follows from an application of results from Hu (2008), and follows two main steps. Before presenting it, we note that, despite its simplicity, this model is not straightforward to estimate: given data on subjects' choices and rewards, we need to estimate choice probabilities conditional on subjects' beliefs, even though these beliefs are not only unobserved, but also changing over time.

**Step one: identification of choice probabilities $\mathbf{P(Y_t|X_t^*)}$ and measurement probabilities $\mathbf{P(Z_t|X_t^*)}$.** We begin with the following factorization:

$$\begin{aligned}
f(Z_t, Y_t|Z_{t-1}) &= \sum_{X_t^*} f(Z_t, Y_t, X_t^*|Z_{t-1}) \\
&= \sum_{X_t^*} f(Z_t|Y_t, X_t^*, Z_{t-1}) f(Y_t, X_t^*|Z_{t-1}) \\
&= \sum_{X_t^*} f(Z_t|Y_t, X_t^*, Z_{t-1}) f(Y_t|X_t^*, Z_{t-1}) f(X_t^*|Z_{t-1}) \\
&= \sum_{X_t^*} f(Z_t|X_t^*) f(Y_t|X_t^*) f(X_t^*|Z_{t-1})
\end{aligned}$$

where the last equality applies assumptions 1 and 3.

For any fixed $Y_t = y$, then, we can write the above in matrix notation as:

$$\mathbf{A}_{y,Z_t|Z_{t-1}} = \mathbf{B}_{Z_t|X_t^*} \mathbf{D}_{y|X_t^*} \mathbf{C}_{X_t^*|Z_{t-1}}$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are all $K \times K$ matrices, and $\mathbf{D}$ is a $K \times K$ diagonal matrix. These are defined as:

$$\begin{aligned}
\mathbf{A}_{y,Z_t|Z_{t-1}} &= \left[ f_{Y_t,Z_t|Z_{t-1}}(y, i|j) \right]_{i,j} \\
\mathbf{B}_{Z_t|X_t^*} &= \left[ f_{Z_t|X_t^*}(i|k) \right]_{i,k} \\
\mathbf{C}_{X_t^*|Z_{t-1}} &= \left[ f_{X_t^*|Z_{t-1}}(k|j) \right]_{k,j} \\
\mathbf{D}_{y|X_t^*} &= \begin{bmatrix} f_{Y_t|X_t^*}(y|1) & 0 & 0 \\ 0 & f_{Y_t|X_t^*}(y|2) & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & f_{Y_t|X_t^*}(y|K) \end{bmatrix}
\end{aligned} \tag{4}$$

Similarly to the above, we can derive that

$$\mathbf{G}_{Z_t|Z_{t-1}} = \mathbf{B}_{Z_t|X_t^*} \mathbf{C}_{X_t^*|Z_{t-1}}$$

where $\mathbf{G}$ is likewise a $K \times K$ matrix, defined as

$$\mathbf{G}_{Z_t|Z_{t-1}} = \left[ f_{Z_t|Z_{t-1}}(i|j) \right]_{i,j}. \tag{5}$$

From Assumption 3(i), we combine the two previous matrix equalities to obtain

$$\mathbf{A}_{y,Z_t|Z_{t-1}} \mathbf{G}_{Z_t|Z_{t-1}}^{-1} = \mathbf{B}_{Z_t|X_t^*} \mathbf{D}_{y|X_t^*} \mathbf{B}_{Z_t|X_t^*}^{-1}. \tag{6}$$

This is an eigenvalue decomposition of the matrix $\mathbf{A}_{y,Z_t|Z_{t-1}}\mathbf{G}^{-1}_{Z_t|Z_{t-1}}$, which can be computed from the observed data sequence $\{Y_t, Z_t\}$.[10] This shows that from the observed data, we can identify the matrices $\mathbf{B}_{Z_t|X^*_t}$ and $\mathbf{D}_{y|X^*_t}$, which are the matrices with entries equal to (respectively) the measurement probabilities $P(Z_t|X^*_t)$ and choice probabilities $P(Y_t|X^*_t)$.

In order for this identification argument to be valid, the eigendecomposition in Eq. (6) must be unique. This requires the eigenvalues in this decomposition (corresponding to choice probabilities $P(y|X^*_t))$ to be distinctive; that is, $P(y|X^*_t)$ should vary in $X^*_t$. This is ensured by Assumption 1. Furthermore, even if the eigendecomposition is unique, the representation in Eq. (6) is invariant to the ordering (or permutation) and scalar normalization of eigenvectors. Assumption 3(ii) imposes the correct ordering on the eigenvectors: specifically, it implies that columns with higher average value correspond to larger value of $X^*_t$. Finally, because the eigenvectors in the decomposition correspond to the conditional probabilities $P(Z_t|X^*_t)$, it is appropriate to normalize each column so that it sums to one. Hence, the uniqueness of the eigendecomposition, coupled with the ordering and normalization assumptions, ensure that the choice probabilities, measurement probabilities, and learning rules can be uniquely identified from the observed matrices $\mathbf{A}$ and $\mathbf{G}$.

**Step two: identification of learning rule probabilities $\mathbf{P(X^*_{t+1}|X^*_t, R_t, Y_t)}$.** Again, start with a factorization

$$\begin{aligned}
&f(Z_{t+1}, Y_t, R_t, Z_t) \\
&= \sum_{X^*_t}\sum_{X^*_{t+1}} f(Z_{t+1}, X^*_{t+1}, Y_t, X^*_t, R_t, Z_t) \\
&= \sum_{X^*_t}\sum_{X^*_{t+1}} f(Z_{t+1}|X^*_{t+1}) f(X^*_{t+1}|Y_t, X^*_t, R_t) f(Z_t|X^*_t) f\,(Y_t, X^*_t, R_t) \\
&= \sum_{X^*_t}\sum_{X^*_{t+1}} f(Z_{t+1}|X^*_{t+1}) f(X^*_{t+1}, Y_t, X^*_t, R_t) f(Z_t|X^*_t)
\end{aligned}$$

where the second equality applies assumptions 1, 2, and 3. Then, for any fixed $Y_t = y$ and $R_t = r$, we have the matrix equality

$$\mathbf{H}_{Z_{t+1},y,r,Z_t} = \mathbf{B}_{Z_{t+1}|X^*_{t+1}}\mathbf{L}_{X^*_{t+1},X^*_t,y,r}\mathbf{B}^T_{Z_t|X^*_t}.$$

[10] From Eq. (5), the invertibility of $\mathbf{G}$ (which is Assumption 3(i)) implies the invertibility of $\mathbf{B}$.

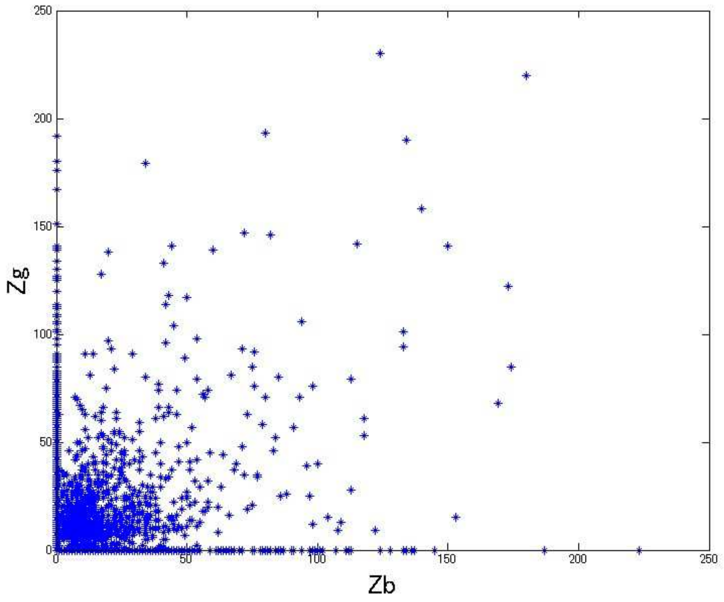


Figure 2: Scatter plot of $Z_b$ (fixation on blue) and $Z_g$(fixation on green)
Both $Z_b$ and $Z_g$ are reported in $2 \times 10^{-2}$ seconds.

The matrices $\mathbf{H}$ and $\mathbf{L}$ are $K \times K$ matrices defined as

$$\begin{aligned} \mathbf{H}_{Z_{t+1},y,r,Z_t} &= \left[f_{Z_{t+1},Y_t,R_t,Z_t}(i,y,r,j)\right]_{i,j} \\ \mathbf{L}_{X^*_{t+1},X^*_t,y,r} &= \left[f_{X^*_{t+1},X^*_t,Y_t,R_t}(i,j,y,r)\right]_{i,j}. \end{aligned} \tag{7}$$

Assumption 4 ensures that $\mathbf{B}_{Z_{t+1}|X^*_{t+1}} = \mathbf{B}_{Z_t|X^*_t}$. Hence, we can obtain $\mathbf{L}_{X^*_{t+1},X^*_t,y,r}$ (corresponding to the learning rule probabilities) directly from

$$\mathbf{L}_{X^*_{t+1},X^*_t,y,r} = \mathbf{B}^{-1}_{Z_{t+1}|X^*_{t+1}} \mathbf{H}_{Z_{t+1},y,r,Z_t} \mathbf{B}^{T,-1}_{Z_t|X^*_t}. \tag{8}$$

This result implies that two periods of data $(Z_t, Y_t, R_t), (Z_{t-1}, Y_{t-1}, R_{t-1})$ are sufficient to identify and estimate this learning model.

## 3.2 Remark on eye-tracking measure

Before proceeding to the estimation results, we discuss the eye-tracking measure $Z$, and present some evidence showing that it is a plausible noisy measure of subjects' beliefs.

Recently, eye-tracking has been employed in economics studies to investigate "what people actually think" in various decision environments. Researchers have used this technology to determine how subjects detect truth-telling or deception in sender-receiver games (Wang, Spezio, and Camerer (forthcoming)), how consumers evaluate comparatively a huge number of commodities, as in a supermarket setting (Reutskaja, Nagel, Camerer, and Rangel (forthcoming)), and the relationship between visual attention (as measured by eye-fixations) and valuation of commodities in choice tasks (cf. Krajbich, Armel, and Rangel (2007), Armel and Rangel (2008), Armel, Beaumel, and Rangel (2008), Rangel (2008)). Specifically, Armel and Rangel (2008) construct a plausible behavioral-neuroscientific model of value computation through visual attentions which sucessfully explains the observed relationship between fixation times and subjects' valuations in their experiments.

Here we use subjects' fixation durations as noisy measures of their beliefs (or valuations) for each slot machine. Let $Z_{pt}$ denote the undiscretized eye-movement measure, and $Z_t$ the discretized measure. Now $Z_{pt}$ is defined as,

$$Z_{pt} = \frac{(Z_{bt} - Z_{gt})}{RT_t}; \tag{9}$$

that is, $Z_{b(g)t}$ is the fixation duration at the blue (green) slot machine, and $RT_t$ is the reaction time, ie. the time between the onset of the trial after fixation, and the subject's choice. (Furthermore, in order to control for subject-specific heterogeneity, we normalize $Z_{pt}$ across subjects by dividing by the subject-specific standard deviation of $Z_{pt}$, across all rounds for each subject.)

Thus, $Z_{pt}$ measures how much longer a subject looks at the blue slot machine than the green one during the $t$-th trial, with a larger (smaller) value of $Z_{pt}$ implying longer fixation time at the blue (green) slot machine. Figure 2 contains the scatter plot of $Z_{bt}$ versus $Z_{gt}$, and Figure 3 is the histogram of $Z_{pt}$. The symmetric distribution around the 45-degree line in Figure 2, along with the symmetric shape around zero in Figure 3, indicates that there is no bias toward a certain color. Also, in the existing literature, it is often reported that human subjects exhibit a "right side bias", tending to gaze towards the right side more frequently. However, our experimental data contains no significant evidence of such a bias.

Moreover, this measure of $Z_{pt}$ is well correlated with actual slot machine choices. Table 1 shows the summary statistics of $Z_p$. The correlation between $Y_t$ (which =2(1) if blue(green) is chosen) and $Z_{pt}$ is 0.7647, which suggests that in this choice setting, a longer fixation duration at an alternative implies a larger probability of choosing it. This also provides some indirect evidence favoring Assumption 3, which posits a monotonic relationship between $Z_t$ and the unobserved beliefs $X_t^*$.

Moreover, from Table 1, we also see that, across all subjects and trials, the high reward is

Table 1: Summary statistics for $Y$, $R$, $Z_p$, $RT$, $Z$

$Y$: subjects' choices
$R$: subjects' rewards
$Z_p$: fixation measure (as defined in Eq. (9))
$RT$: reaction time (in $10^{-2}$ seconds)
$Z$: discretized version of $Z_p$

| | green | blue |
|---|---|---|
| $Y$ | 2108 | 2092 |

| | win (\$0.50) | lose (-\$0.50) |
|---|---|---|
| $R$ | 2398 | 1802 |

| | mean | median | upper 5% | lower 5% |
|---|---|---|---|---|
| $Z_p$ | -0.0309 | 0 | 1.3987 | -1.4091 |
| $RT$ | 88.22 | 59.3 | 212.2 | 36.8 |

| Sample size | 21 subjects | 168 blocks | 4200 trials |
|---|---|---|---|
| Corr.($Y$,$Z_p$) | | | 0.7647 |

$Z$ (after two-value discretization)$^A$

| 1(green, $Z_p < 0$) | 2(blue, $Z_p \geq 0$) |
|---|---|
| 2032 | 2168 |

$Z$ (after three-value discretization)$^A$

| 1(green) | 2(not sure) | 3(blue) |
|---|---|---|
| 1887 | 540 | 1773 |

$^A$: for more details on discretizing $Z$, see the appendix, section B obtained with frequency of roughly 57% ($\approx 2398/(2398+1802)$). This is slightly higher than, but significantly different from, 55%, which is the frequency which would obtain if the subjects were choosing completely randomly.[11]

## 4 Estimation

For the estimation, we assume that the variables $Z_t$ and $X_t^*$ are discrete, and take either two or three values. Since the eye-movement measure $Z_t$ is continuous, we must discretize it for estimation. We leave the details of our discretization procedure in Appendix B.

Our estimation procedure mimicks the two-step identification argument from the previous section. That is, for fixed values of $(y, r)$, we first form the matrices **A**, **G**, and **H** (as defined previously) from the observed data, using sample frequencies to estimate the corresponding probabilities. Then we obtain the matrices **B**, **D**, and **L** using the matrix manipulations in Eqs. (6) and (8).

One technical feature is that, because all the elements in the matrices of interest **B**, **D**, and **L** correspond to probabilities, they must take values within the unit interval. However, in the actual estimation, we found that occasionally the estimates do go outside this range. In these cases, we obtained the estimates by a least-squares fitting procedure, where we minimized the elementwise sum-of-squares corresponding to Eqs. (6) and (8), and explicitly restricted each element of the matrices to lie $\in [0,1]$. This was not a frequent recourse; only a handful of the estimates reported below needed to be restricted in this manner.

In addition, while the identification argument above was "cross-sectional" in nature, being based upon two observations of $\{Y_t, Z_t, R_t\}$ per subject, in the estimation we exploited the long time series data we have for each subject, and pooled every two time-contiguous observations $\{Y_{i,r,\tau}, Z_{i,r,\tau}, R_{i,r,\tau}\}_{\tau=t-1}^{\tau=t}$ across all subjects $i$, all blocks $r$, and all trials $\tau = 2, \ldots, 25$. Formally, this is justified under the assumption that the process $\{Y_t, Z_t, R_t\}$ is stationary and ergodic for each subject and each block; under these assumptions, the ergodic theorem ensures that the (across time and subjects) sample frequencies used to construct the matrices **A**, **G**, and **H** converge towards population counterparts.[12]

[11] This is the marginal probability of a good reward, which equals $0.5(0.7+0.4)$ from Eq. (1). The t-statistic for the null that subjects are choosing randomly equals 169.67, so that hypothesis is strongly rejected.x

[12] While the results reported below were obtained by pooling the data across all subjects, we also estimated the model separately for the subsamples of Caltech students, vs. community members. There were few

Before presenting the results, we present some Monte Carlo simulation results in Table 2, for simulated datasets around the same size as the datasets drawn from our experiments. These show that the estimation procedure produces accurate estimates of the model components, with the differences between the estimated and actual values usually on the order of magnitude of $10^{-1}$ times the parameter value.

## 4.1 Estimation results

### 4.1.1 Two-value estimates

In Table 3, we present estimates in the specification where $X_t^*$ and $Z_t$ are assumed to be binary variables taking values $\in \{1, 2\}$. The standard errors, shown in parentheses, were computed using block bootstrap resampling (using 1000 iterations, resampled from all 168 blocks).

Starting from the top of the table, we see that the choice probabilities are reasonable, and very much aligned with beliefs. When $X_t^* = 1$ (associated with beliefs that "green is currently the good state"), then the green slot machine is pulled 98% of the time. Similarly, when $X_t^* = 2$, then the blue slot machine is chosen 94% of the time. In many learning settings (including reinforcement learning, cf. Sutton and Barto (1998, pg. 28), as well as Bayesian learning), an optimal decision rule require choices to not be completely in line with current beleifs; to avoid getting "stuck" at suboptimal choices, subjects should explore with some small probability. However, as we noted before (cf. remark 1), this incentive for exploration is reduced in our reversal learning experiment, and so the small estimate of $\epsilon$ here is reasonable.

**Remark 3 (What do the beliefs $\{X_t^*\}$ mean?)** *As we discussed earlier in Remark 1, in the standard multi-armed bandit model, subjects' choices of which arm to pull depends on the dynamic allocation, or "Gittins" index, which depends not only on current beliefs about which arm yields a higher return, but also on the informational value in pulling an arm which may not be currently optimal, but which may yield information useful in future decisions. However, in our reversal learning setting, because the returns in the two arms are negatively correlated, this informational value term is largely nonexistent. Therefore, in the context of such a model, we can quite confidently interpret the unobserved variables $X_t^*$, which completely determine subjects' choices in our learning model, as a measurement of subjects'*

noticeable differences in the results across these classes of subjects.

Table 2: Monte Carlo Results. (2500 iterations, median, “”= true value)

Each cell contains the median parameter value across all iterations, and the actual parameter value in double quotes. Standard deviations across all iterations are in parentheses. Note that columns sum to one.

$P(Y_t|X_t^*)$

| $X_t^*$ | 1(green) | 2(blue) |
|---|---|---|
| $Y_t = 1$ | 0.9502 | 0.0500 |
| (green) | “0.9500” | “0.0500” |
| | (0.0250) | (0.0245) |
| 2 | 0.0498 | 0.9500 |
| (blue) | ”0.0500” | ”0.9500” |

$P(Z_t|X_t^*)$

| $X_t^*$ | 1(green) | 2(blue) |
|---|---|---|
| $Z_t = 1$ | 0.9002 | 0.1002 |
| (green) | “0.9000” | “0.1000” |
| | (0.0221) | (0.0228) |
| 2 | 0.0998 | 0.8998 |
| (blue) | “0.1000” | “0.9000” |

$P(X_{t+1}^*|X_t^*, y, r)$, r = 1(lose), y = 1(green)

| $X_t^*$ | 1(green) | 2(blue) |
|---|---|---|
| $X_{t+1}^* = 1$ | 0.3997 | 0.1782 |
| (green) | “0.4000” | “0.1500” |
| | (0.0314) | (0.1959) |
| 2 | 0.6003 | 0.8218 |
| (blue) | “0.6000” | “0.8500” |

$P(X_{t+1}^*|X_t^*, y, r)$, r = 2(win), y = 1(green)

| $X_t^*$ | 1(green) | 2(blue) |
|---|---|---|
| $X_{t+1}^* = 1$ | 0.8002 | 0.7073 |
| (green) | “0.8000” | “0.7000” |
| | (0.0283) | (0.2031) |
| 2 | 0.1998 | 0.2927 |
| (blue) | “0.2000” | “0.3000” |

Note: Learning rules for y = 2(blue) is practically the same as for y = 1(green), so we omit them for the sake of brevity.

*current beliefs regarding which arm is currently the "good" one. Thus another benefit of a reversal learning model is the unambiguity in interpreting the unobserved "beliefs" $X_t^*$ in this setting.*

The second panel in Table 3 contains the measurement probabilities. The estimates imply that beliefs closely track the eye-movement measures, with (for instance) beliefs favoring green leading to longer gazes at the green slot machine on the computer screen around 92% of the time.

Finally, the remaining panels present the learning rule probabilities for all four configurations of $(R_t, Y_t) \in \{(1,1),(2,1),(1,2),(2,2)\}$. The columns and rows are ordered differently across the panels, for ease of interpreting the results. Generally, the left column of each panel makes sense. Comparing the third and fourth panels in Table 3, we see that given the choice of "green" ($Y_t = 1$) and given beliefs in favor of green ($X_t^* = 1$), a higher reward leads to more intense updating of beliefs towards green in the next trial; that is:

$$\begin{aligned} 0.87 &= P(X_{t+1}^* = 1 | X_t^* = 1, R_t = 2, Y_t = 1) \\ >>&P(X_{t+1}^* = 1 | X_t^* = 1, R_t = 1, Y_t = 1) = 0.54. \end{aligned}$$

Similarly, comparing the bottom two panels, we see that if the subject is predisposed towards blue ($X_t^* = 2$) then choosing blue $Y_t = 2$ and obtaining the higher reward $R_t = 2$ leads subjects to place a belief of 90% on "blue" the following trial, vs. only 54% if this led to the lower reward $R_t = 1$.

On the other hand, the right columns in these panels are a bit puzzling. They indicate a great deal of state dependence in beliefs, when one chooses actions which are contrary to beliefs. For example, the third and fourth panels indicate that when $X_t^* = 2$ (so current beliefs favor "blue"), but the subject chooses $Y_t = 1$ ("green"), then the updated beliefs are not affected much by the reward: with a high reward, beliefs switch to "green" ($X_{t+1}^* = 1$) with only 25% probability, but with a low reward, beliefs switched to "green" with the *slightly higher* probability of 30%, which is puzzling. Similarly, in the bottom two panels, when current beliefs favor "green" ($X_t^* = 1$), but the blue slot machine was chosen ($Y_t = 2$), then the probability that beliefs switched to "blue" ($X_{t+1}^* = 2$) is slightly higher following a low rather than high reward.

At face value, this suggests that subjects do not update their beliefs properly following "exploratory" (ie. contrary to belief) actions. However, as we will see now, these puzzling results are no longer so apparent when we allow beliefs to take three distinct values.

Table 3: Two-value estimates: Specification where $X_t^*$ and $Z_t$ are binary

Each cell contains parameter estimates, with bootstrapped standard errors in parentheses. Note that each column sums to one.

$P(Y_t|X_t^*)$

| $X_t^*$ | 1(green) | 2(blue) |
|---|---|---|
| $Y_t = 1$<br>(green) | 0.9756<br>(0.0115) | 0.0573<br>(0.0165) |
| 2<br>(blue) | 0.0244 | 0.9427 |

$P(Z_t|X_t^*)$

| $X_t^*$ | 1(green) | 2(blue) |
|---|---|---|
| $Z_t = 1$<br>(green) | 0.9093<br>(0.0156) | 0.0888<br>(0.0116) |
| 2<br>(blue) | 0.0907 | 0.9112 |

$P(X_{t+1}^*|X_t^*, y, r)$, r = 1(lose), y = 1(green)

| $X_t^*$ | 1(green) | 2(blue) |
|---|---|---|
| $X_{t+1}^* = 1$<br>(green) | 0.5401<br>(0.0279) | 0.2950<br>(0.1588) |
| 2<br>(blue) | 0.4599 | 0.7050 |

$P(X_{t+1}^*|X_t^*, y, r)$, r = 2(win), y = 1(green)

| $X_t^*$ | 1(green) | 2(blue) |
|---|---|---|
| $X_{t+1}^* = 1$<br>(green) | 0.8695<br>(0.0256) | 0.2471<br>(0.2160) |
| 2<br>(blue) | 0.1305 | 0.7529 |

$P(X_{t+1}^*|X_t^*, y, r)$, r = 1(lose), y = 2(blue)

| $X_t^*$ | 2(blue) | 1(green) |
|---|---|---|
| $X_{t+1}^* = 2$<br>(blue) | 0.5407<br>(0.0263) | 0.6836<br>(0.2249) |
| 1<br>(green) | 0.4593 | 0.3164 |

$P(X_{t+1}^*|X_t^*, y, r)$, r = 2(win), y = 2(blue)

| $X_t^*$ | 2(blue) | 1(green) |
|---|---|---|
| $X_{t+1}^* = 2$<br>(blue) | 0.9003<br>(0.0242) | 0.6146<br>(0.2287) |
| 1<br>(green) | 0.0997 | 0.3854 |

### 4.1.2 Three-value estimates

Tables 4 and 5 present results from a specification where $X_t^*$ is assumed to take three values $\{1, 2, 3\}$, and likewise $Z_t$ is discretized to take these three values. We interpret $X^* = 1, 3$ as indicative of "strong beliefs" favoring (respectively) green and blue, while the intermediate value $X^* = 2$ indicates that the subject is "not sure".

Table 4 contains the estimates of the choice and measurement probabilities.[13] The first and last columns of the panels in this table indicate that choices and eyes movements are closely aligned with beliefs, when beliefs are sufficiently strong (ie. are equal to either $X^* = 1$ or $X^* = 3$). Specifically, in these results, the "exploration probability" is smaller than in the two-value results, being equal to 1.3% when $X_t^* = 1$, and only 0.64% when $X_t^* = 3$. As we remarked above, such small probabilities can be consistent with optimal behavior, in our reversal learning environment.

When $X_t^* = 2$, however, suggesting that the subject is unsure of the state, there is a slight bias in choices towards "blue", with $Y_t = 2$ roughly 56% of the time. The bottom panel indicates that when subjects are not sure, they tend to gaze in the middle of the screen, around 63% of the time.

The learning rule estimates are presented in Table 5. The results are similar to the two-value results, but some of the problems from those results disappear when we allow beliefs to take three values. The left columns show how beliefs are updated when "exploitative" choices (ie. choices made in accordance with beliefs) are taken. We see that when current beliefs indicate "green" ($X_1^* = 1$) and green is chosen ($Y_t = 1$), beliefs are quite responsive to the reward: if $R_t = 1$ (the low reward), then beliefs stay at green with probability 57%, but if $R_t = 2$ (high reward), then this probability is much higher, at 89%. On the other hand, even after positive (ie. high reward) exploitative choices, beliefs may still update towards "blue" ($X_{t+1}^* = 3$) with an 11% chance, rather than sticking at the intermediate level $X_{t+1}^* = 2$. This non-smooth "extremal" updating is a distinctive feature of our learning rule estimates, and is consistent with optimal belief-updating in a probabilistic reversal context: even if the subject were completely sure that "green" after a high reward, she still must consider the

[13] We also considered a robustness check against the possibility that subjects' fixations immediately before making their choices coincide exactly with their choice. While this is not likely in our experimental setting, because subjects were required to indicate their choice by pressing a key on the keyboard, rather than clicking on the screen using a mouse, we nevertheless re-estimated the models but eliminating the last segment of the reaction time in computing the $Z_t$. The results are very similar to the reported results, both qualitatively and quantitatively.

Table 4: Three-value estimates: Specification where $X_t^*$ and $Z_t$ take three values

Each cell contains parameter estimates, with bootstrapped standard errors in parentheses. Note that each column sums to one.

Choice probabilities:

$P(Y_t|X_t^*)$

| $X_t^*$ | 1(green) | 2(not sure) | 3(blue) |
|---|---|---|---|
| $Y_t = 1$<br>(green) | 0.9866<br>(0.0561) | 0.4421<br>(0.1274) | 0.0064<br>(0.0146) |
| 2<br>(blue) | 0.0134 | 0.5579 | 0.9936 |

$P(Z_t|X_t^*)$

| $X_t^*$ | 1(green) | 2(not sure) | 3(blue) |
|---|---|---|---|
| $Z_t = 1$<br>(green) | 0.8639<br>(0.0468) | 0.2189<br>(0.1039) | 0.0599<br>(0.0218) |
| 2<br>(middle) | 0.0815<br>(0.0972) | 0.6311<br>(0.1410) | 0.0980<br>(0.0369) |
| 3<br>(blue) | 0.0546<br>(0.0581) | 0.1499<br>(0.1206) | 0.8421<br>(0.0529) |

Table 5: Three-value estimates: Specification where $X^*_t$ and $Z_t$ take three values

Each cell contains parameter estimates, with bootstrapped standard errors in parentheses. Note that each column sums to one.

Learning Rule updating probabilities:

$P(X^*_{t+1}|X^*_t, y, r)$, r = 1(lose), y = 1(green)

| $X^*_t$ | 1(green) | 2 (not sure) | 3(blue) |
|---:|---:|---:|---:|
| $X^*_{t+1} = 1$<br>(green) | 0.5724<br>(0.0694) | 0.3075<br>(0.0881) | 0.1779<br>(0.2257) |
| 2<br>(not sure) | 0.0000<br>(0.0662) | 0.3138<br>(0.1042) | 0.4002<br>(0.2284) |
| 3<br>(blue) | 0.4276<br>(0.0624) | 0.3787<br>(0.0945) | 0.4219<br>(0.2195) |

$P(X^*_{t+1}|X^*_t, y, r)$, r = 2(win), y = 1(green)

| $X^*_t$ | 1(green) | 2 (not sure) | 3(blue) |
|---:|---:|---:|---:|
| $X^*_{t+1} = 1$<br>(green) | 0.8889<br>(0.0894) | 0.6621<br>(0.1309) | 0.8242<br>(0.2734) |
| 2<br>(not sure) | 0.0000<br>(0.0911) | 0.2702<br>(0.1297) | 0.1758<br>(0.1981) |
| 3<br>(blue) | 0.1111<br>(0.0340) | 0.0678<br>(0.0485) | 0.0000<br>(0.1876) |

$P(X^*_{t+1}|X^*_t, y, r)$, r = 1(lose), y = 2(blue)

| $X^*_t$ | 3(blue) | 2 (not sure) | 1(green) |
|---:|---:|---:|---:|
| $X^*_{t+1} = 3$<br>(blue) | 0.5376<br>(0.0890) | 0.2297<br>(0.0731) | 0.2123<br>(0.1436) |
| 2<br>(not sure) | 0.0458<br>(0.0732) | 0.2096<br>(0.0958) | 0.1086<br>(0.1524) |
| 1<br>(green) | 0.4166<br>(0.0874) | 0.5607<br>(0.0968) | 0.6792<br>(0.1881) |

$P(X^*_{t+1}|X^*_t, y, r)$, r = 2(win), y = 2(blue)

| $X^*_t$ | 3(blue) | 2 (not sure) | 1(green) |
|---:|---:|---:|---:|
| $X^*_{t+1} = 3$<br>(blue) | 0.8845<br>(0.1000) | 0.6163<br>(0.1136) | 0.6319<br>(0.1647) |
| 2<br>(not sure) | 0.0000<br>(0.0968) | 0.3558<br>(0.1160) | 0.3566<br>(0.1637) |
| 1<br>(green) | 0.1155<br>(0.0499) | 0.0279<br>(0.0373) | 0.0116<br>(0.0679) |

possibility that the good state could change to "blue" by the next trial, due to stochastic state process.

The results in the right-most columns, describing belief updating following "explorative" (contrarian) choices, are on the whole more sensible than in the two-value estimates. For instance, considering the top two panels, when current beliefs are favorable to "blue" ($X^*_t = 3$), but "green" is chosen, beliefs update more towards "green" ($X^*_{t+1} = 1$) after a low rather than high reward (82% vs. 18%)

The second columns in these panels show how beliefs evolve following (almost-) random choices. Again considering the top two panels, we see that when current beliefs are unsure ($X^*_t = 2$), there is stronger updating towards "green" when green choice yielded the higher reward (66% vs. 31%). The results in the bottom two panels are very similar to those in the top two panels, but describes how subjects update beliefs following choices of "blue" ($Y_t = 2$).

## 5 Comparing nonparametric vs. standard learning models

In this section, we compare the beliefs implied by our estimated learning model (which we will refer to as the "nonparametric" model, for convenience), to those implied by alternative learning models. We consider two alternative parametric learning rules: Bayesian and reinforcement learning. Given that our learning rule was estimated nonparametrically, and in that sense encompasses the other two models, we examine which of these two popular alternative models is closer to our nonparametric learning model. Appendix A contains additional details on how the beliefs were derived for each of these three learning models.

Figures 3-5 contains the raw histograms for the (noisy) measurements of beliefs from the three learning models: Figure 3 contains the histogram of the eye tracking measure $Z$, which is used to pin down beliefs in our nonparametric learning model. Figure 4 contains the histogram of the Bayesian posterior probabilities, computed given our experimental design and the observed data. Finally, Figure 5 contains the histogram for the difference in the calibrated valuation measures for the "blue" vs. "green" slot machine, from a TD-learning reinforcement learning model.

A noteworthy feature is that the histograms for the eye-tracking measure $Z$ and the TD-learning valuations look similar: both are trimodal. The Bayesian posterior mean measure, on the other hand, is unimodel. As we will see later, this implies that beliefs from the

Figure 3: Histogram of $Z_p = \frac{Z_b - Z_g}{RT}$

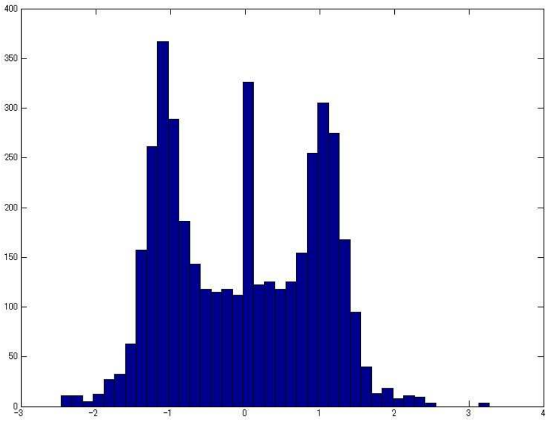


Figure 4: Histogram of Bayesian Belief $B^*$

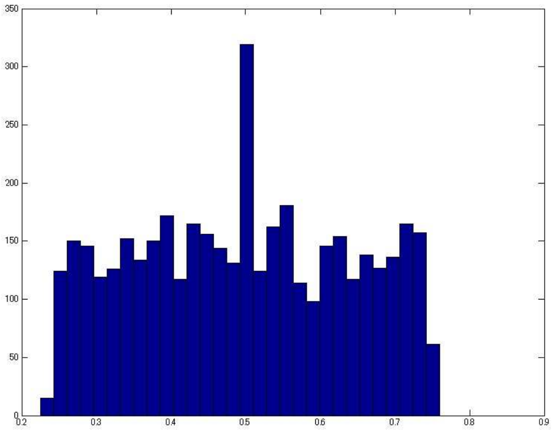
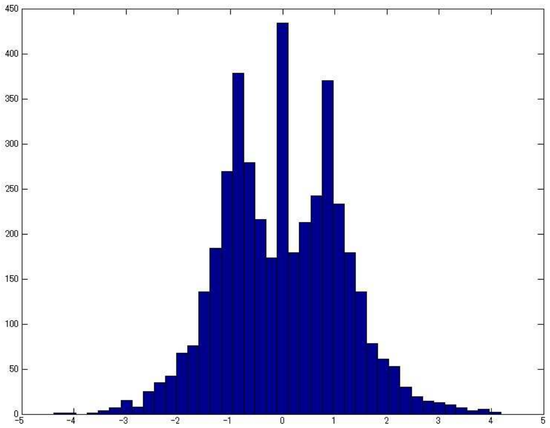


Figure 5: Histogram of $V^* = V_b - V_g$ in RL



nonparametric model will be closer to the RL model, than the Bayesian model. Moreover, we will also see that the Bayesian learning model tends to predict "smoother" choice behavior than what we observe in the data, while the beliefs from the nonparametric model are "jumpy" in comparison.

**Overall summary statistics** In Table 6, we present some summary statistics which describe the predictive success of our nonparametric learning model (as given by the optimally-fitted beliefs $X_t^*$), vs. the Bayesian beliefs $B^*$ and the valuations $V^*$ in the RL learning model. For simplicity, we will abuse terminology somewhat and refer in what follows to $X^*$, $V^*$, and $B^*$ as the "beliefs" implied by, respectively, our nonparametric model, the RL model, and the Bayesian model. This table contains eight panels.

Panel 1 gives the total tally, across all subjects, blocks, and trials, of the number of times the nonparametric beliefs $X^*$ took each of the three values. We see that subjects' beliefs tended to favor green and blue toughly equally, with "not sure" lagging far behind. The close split between "green" and "blue" beliefs is consistent with the notion that subjects have rational expectations, with flat priors on the unobserved state $S_1$ at the beginning of each block. The second panel shows analogous statistics for the beliefs from the RL and Bayesian models. The RL valuation measure $V^*$ appears largely symmetric and centered around zero, while the average Bayesian $B^*$ lies also around 0.5. Thus, on the whole, all three measures of beliefs appear equally distributed between "green" and "blue".

Panel 3 contains the pairwise correlation among ($X^*$, $V^*$, $B^*$), the beliefs from the three models. The correlation between $X^*$ and $V^*$ (0.59) exceeds that between $X^*$ and $B^*$ (0.53). This shows that the nonparametric beliefs $X^*$ are, stochastically, more similar to the RL beliefs $V^*$ than to the Bayesian beliefs $B^*$; it also confirms the evidence from the histograms, as described above. The Bayesian model is the most restrictive one, and imposes the highest degree of rationality on subjects, which may explain its inferior fit, relative to the RL model, to our nonparametric learning rule. However, the correlations between our nonparametric beliefs $X^*$ and $B^*$ and $V^*$ are markedly lower than that between $B^*$ and $V^*$ (which is 0.82). This indicates that, informationally, the beliefs from the Bayesian and RL models are very similar.

The next panel shows that the correlation of $X^*$ with the observed choices $Y$ is much higher (0.7552) than the correlation of choices with the other measures. This superior performance of the nonparametric beliefs in predicting subjects' choices is not too surprising, since the beliefs are estimated from the data, whereas the other two models are only calibrated to the

Table 6: Summary statistics for the three models

**Panel 1**:

| $X^*$ | 1(green) | 2(not sure) | 3(blue) |
|---|---|---|---|
| | 1878 | 366 | 1956 |

**Panel 2**:

| | mean | median | std. | 1/3 quantile | 2/3 quantile |
|---|---|---|---|---|---|
| $B^*$ (Bayesian Belief) | 0.4960 | 0.5000 | 0.1433 | 0.4201 | 0.5644 |
| $V^*(= V_b - V_g)$ | -0.0035 | 0 | 1.1152 | -0.6588 | 0.6068 |

**Panel 3**: Correlations in the three models

| | |
|---|---|
| Corr.$(X^*, V^*)$ | 0.5874 (0.0014)* |
| Corr.$(X^*, B^*)$ | 0.5274 (0.0013) |
| Corr.$(B^*, V^*)$ | 0.8271 (0.0006) |

*: bootstrapped standard error in parentheses

**Panel 4**: Correlations with observed choices $Y$ (all samples)

| | |
|---|---|
| Corr.$(Y, X^*)$ | 0.7552 |
| Corr.$(Y, V^*)$ | 0.5560 |
| Corr.$(Y, B^*)$ | 0.5175 |

**Panel 5**: Correlations with choices $Y$ (excluding intermediate beliefs)

| | | |
|---|---|---|
| Corr.$(Y, X^*)$ | 0.7906 | (keep only $X^*$ =1,3) |
| Corr.$(Y, V^*)$ | 0.6786 | (keep only $V^* \notin$ [1/3 quant., 2/3 quant.]) |
| Corr.$(Y, B^*)$ | 0.6252 | (keep only $B^* \notin$ [1/3 quant., 2/3 quant.]) |

**Panel 6**: Correlations with choices Y (last 10 trials, first 5 trials)

| | last 10 | first 5 |
|---|---|---|
| Corr.$(Y, X^*)$ | 0.7474 | 0.6908 |
| Corr.$(Y, V^*)$ | 0.5582 | 0.5201 |
| Corr.$(Y, B^*)$ | 0.5267 | 0.4678 |

**Panel 7**: Number of "explorative" (belief non-congruent) choices $Y$

| | |
|---|---|
| Nonparametric | 402 |
| Reinforcement Learning | 455 |
| Bayesian | 543 |

**Panel 8**: Correlations with noisy measure Z (NB: Corr.$(Z, Y) = 0.7738$)

| | |
|---|---|
| Corr.$(Z, X^*)$ | 0.8575 |
| Corr.$(Z, V^*)$ | 0.4717 |
| Corr.$(Z, B^*)$ | 0.4296 |

data. The next two panels break down the correlation between the observed choices and the difference measures of beliefs, for subsamples of the data. Panel 5 only considers subjects' choices when the implied beliefs are strong (in the sense of taking extreme values). For the nonparametric model, we omitted observations when $X^*$ was estimated to be "not sure", while for the other two models, we omitted observations when beliefs lay between the 1/3 and 2/3 quantile. The results show that when beliefs are strong, the nonparametric model continues to predict choices better than the Bayesian and RL models. Panel 6 similarly shows that predicted choice behavior using only the last ten rounds of each subjects' data, or the first five rounds, is more accurate for all three models; but as before, the nonparametric model is more accurate.[14]

The better predictive fit of the nonparametric beliefs $X^*$ implies that our nonparametric model should classify fewer choices as "exploratory" ones (where exploratory behavior is generally defined as making contrarian choices in the face of strong beliefs). This intuition is confirmed in Panel 7, which shows that the nonparametric model classifies only 405 (10.5%) of the subjects' choices as exploratory. The RL model which, as pointed above, is closer to our nonparmetric model, classifies 455 of the choices as exploratory, while the Bayesian model classifies 543 choices as such.

Finally, the bottom panel shows the sample correlation between the eye-movement measure, and the implied beliefs. Not surprisingly, the correlation is much higher for the nonparametric beliefs $X^*$ (since identification of the nonparametric model relies on the monotonicity condition in Assumption 3). The Bayesian and RL beliefs, which do not require $Z$ to compute, exhibit a smaller correlation with $Z$.

**A closer look at individual blocks** To look more closely at the differences between the three learning models, we plot, in Figures 6-9, the actual choices, as well as subjects' beliefs regarding which slot is better, from the three learning models, for four representative subject-blocks of choices. The actual choices are plotted in crosses (+'s), with higher crosses (at 0.25) signifying "blue" and lower crosses (at -0.25) signifying "green". The subject's beliefs from the three models, all recentered and rescaled around zero, are plotted in three

[14] While the predictions using the nonparametric model reported here were "in sample" (that is, the estimation and prediction were done using the same sample), we also considered out of sample prediction (where the estimation and prediction were performed on different subsamples of subjects) and the results were very similar.

different lines.[15]

Figure 6, for trial #4 of subject #6, is typical. Comparing the predicted choices, we see that, generally, all three models perform reasonably well. The choice of “blue” in trial #18 was unanticipated by all three models, and would be classified as “exploratory” in each case. In this block, the Bayesian and RL beliefs move in tandem. Hence, the choice of “green” in trial #8 was a surprise to the nonparametric model, but predicted by the other two models. On the other hand, the choice of “green” in trial #9 was predicted by the nonparametric beliefs, but not by the Bayesian and RL models.

Figure 7, which shows subject (#4) and block (#6), presents an example where the Bayesian and RL beliefs diverge, at the end of the block. It is noteworthy here that the Bayesian model “misses” the final run of “green” choices. On the other hand, the nonparametric and RL beliefs are able to predict these choices. Also note here that when the Bayesian and RL beliefs diverge, then the nonparametric beliefs are closer to the RL beliefs, which was apparent from the summary statistics discussed earlier, which indicated a stronger correlation between the nonparametric and RL beliefs, than between the nonparametric and Bayesian beliefs.

The two remaining figures (8 and 9) contain additional instances of choices which all three learning models would classify as “exploratory”. These are trials #12, #15, and #18 in Figure 8, and trials #11, #13, and #25 in Figure 9. Across all four figures here, the nonparametric beliefs $X^*$ jump between favoring “blue” and “green”, and rarely take the intermediate value “not sure”. This is consistent with the estimates of the learning rule, especially the left-hand side columns of the panels in Table 5, which place zero probability on $X^* = 2$ following choices congruent with current beliefs. Both the Bayesian and RL model posit a smoother belief updating process. This “jumpiness” in the nonparametric learning rule represent another important qualitative difference relative to the standard learning models.

## 6 Conclusions

In this paper, we estimate learning rules nonparametrically from data drawn from experiments of multi-armed bandit problems. The experimental data are augmented by mea-

[15] That is, the Bayesian beliefs were plotted as $B_t^* - 0.5$, while the RL beliefs were plotted as $V_t^* = 0.25 * (V_b^t - V_g^t)$. The nonparametric beliefs were plotted as $0.25 * (X_t^* - 2)$.

Figure 6: Subject 6, block 4

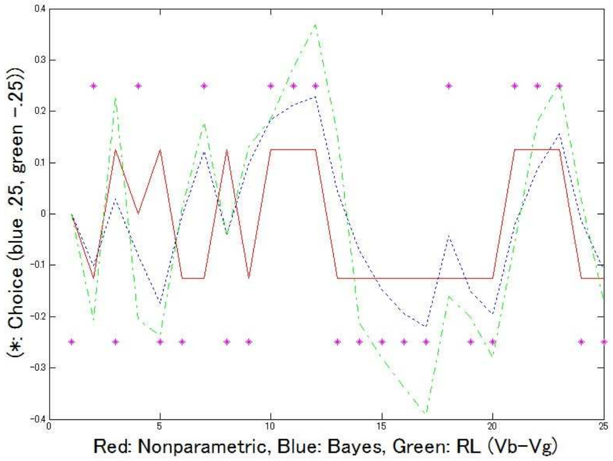


Figure 7: Subject 4, block 6

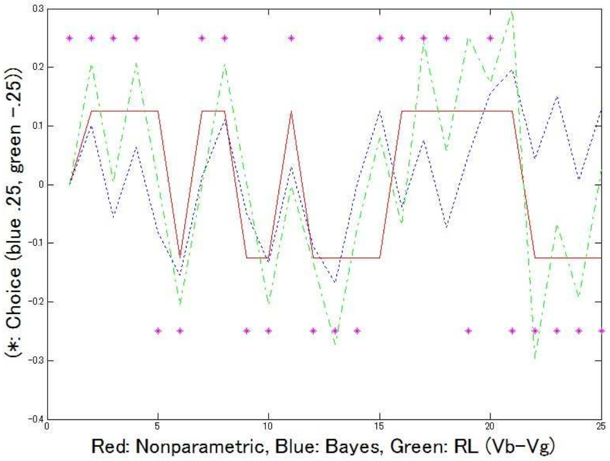


Figure 8: Subject 5, block 8

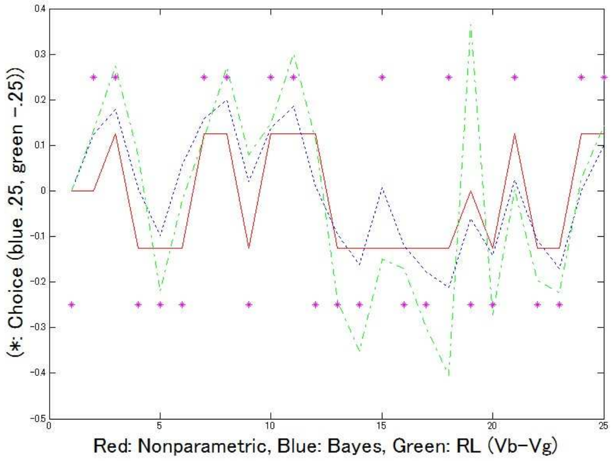
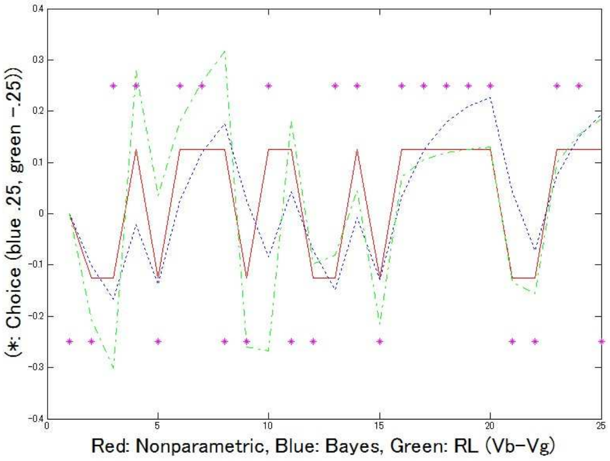


Figure 9: Subject 1, block 3



surements of subjects' eye movements from an eye tracker machine, which play the role of auxiliary measures of subjects' beliefs. Our estimated learning rules have some distinctive features – notably, non-smooth updating following positive "exploitative" choices. A comparison of the nonparametric learning rules with "standard" learning models shows that our estimates are closer to the reinforcement learning model, than to a Bayesian model. Altogether, our analysis points out some deficiencies in the Bayesian model as a descriptive model, thus echoing previous findings in both the experimental and finance literatures.

Our nonparametric estimator for subjects' choice probabilities and learning rules is easy to implement. Potentially, it can also be applied to other experimental settings where auxiliary measures of subjects' beliefs and valuations are available, such as the typical neuroscience fMRI setting.

# References

Ackerberg, D. (2003): "Advertising, Learning, and Consumer Choice in Experience Good Markets: A Structural Examination," *International Economic Review*, 44, 1007–1040.

Armel, K., A. Beaumel, and A. Rangel (2008): "Biasing simple choices by manipulating relative visual attention," *Judgment and Decision Making*, 3(5), 396–403.

Armel, K., and A. Rangel (2008): "The impact of computation time and experience on decision values," *American Economic Review*, 98(2), 163–168.

Banks, J., and R. Sundarum (1992): "Denumerable-Armed Bandits," *Econometrica*, 60, 1071–1096.

Behrens, T., M. Woolrich, M. Walton, and M. Rushworth (2007): "Learning the value of information in an uncertain world," *Nature Neuroscience*, 10(9), 1214–1221.

Boorman, E., T. Behrens, M. Woolrich, and M. Rushworth (2009): "How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action," *Neuron*, 62(5), 733–743.

Chan, T. Y., and B. H. Hamilton (2006): "Learning, Private Information, and the Economic Evaluation of Randomized Experiments," *Journal of Political Economy*, 114, 997–1040.

Charness, G., and D. Levin (2005): "When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect," *American Economic Review*, 95, 1300–1309.

Choi, J., D. Laibson, B. Madrian, and A. Metrick (2009): "Reinforcement learning and savings behavior," *The Journal of Finance*, 64(6), 2515–2534.

Crawford, G., and M. Shum (2005): "Uncertainty and Learning in Pharmaceutical Demand," *Econometrica*, 73, 1137–1174.

Daw, N., J. O'Doherty, P. Dayan, B. Seymour, and R. Dolan (2006): "Cortical substrates for exploratory decisions in humans," *Nature*, 441(7095), 876–879.

Doya, K. (2002): "Metalearning and neuromodulation," *Neural Networks*, 15(4-6), 495–506.

El-Gamal, M., and D. Grether (1995): "Are People Bayesian? Uncovering Behavioral Strategies," *Journal of American Statistical Association*, 90, 1137–1145.

Erdem, T., and M. Keane (1996): "Decision-making Under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets," *Marketing Science*, 15, 1–20.

Ghahramani, Z. (2001): "An Introduction to Hidden Markov Models and Bayesian Networks," *International Journal of Pattern Recognition and Artificial Intelligence*, 15, 9–42.

Gittins, J., and G. Jones (1974): "A Dynamic Allocation Index for the Sequential Design of Experiments," in *Progress in Statistics*, ed. by e. a. J. Gani. North-Holland.

Glimcher, P., C. Camerer, R. Poldrack, and E. Fehr (2008): *Neuroeconomics: decision making and the brain.* Academic Press.

Grether, D. M. (1992): "Testing bayes rule and the representativeness heuristic: Some experimental evidence," *Journal of Economic Behavior & Organization*, 17, 31–57.

Hampton, A., P. Bossaerts, and J. O'Doherty (2006): "The Role of the Ventromedial Prefrontal Cortex in Abstract State-Based Inference during Decision Making in Humans," *Journal of Neuroscience*, 26, 8360–8367.

Hu, Y. (2008): "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: a General Solution," *Journal of Econometrics*, 144, 27–61.

Hu, Y., and M. Shum (2008): "Nonparametric Identification of Dynamic Models with Unobserved State Variables," Jonhs Hopkins University, Dept. of Economics working paper #543.

Imai, S., N. Jain, and A. Ching (2009): "Bayesian Estimation of Dynamic Discrete Choice Models," *Econometrica*, 77, 1865–1899.

Krajbich, I., C. Armel, and A. Rangel (2007): "Visual attention drives the computation of value in goal–directed choice," Discussion paper, Working Paper, Caltech.

Kuhnen, C., and B. Knutson (2008): "The Influence of Affect on Beliefs, Preferences and Financial Decisions," MPRA Paper 10410, University Library of Munich, Germany.

Marcoul, P., and Q. Weninger (2008): "Search and active learning with correlated information: Empirical evidence from mid-Atlantic clam fishermen," *Journal of Economic Dynamics and Control*, 32, 1921–1948.

Odean, T., M. Strahilevitz, and B. Barber (2004): "Once Burned, Twice Shy: How Naive Learning and Counterfactuals Affect the Repurchase of Stocks Previously Sold," Discussion paper, mimeo., UC Berkeley, Haas School.

Pakes, A., and P. McGuire (2001): "Stochastic Algorithms, Symmetric Markov Perfect Equilibrium, and the 'Curse' of Dimensionality," *Econometrica*, 69, 1261–1282.

Payzan, É., and P. Bossaerts (2009): "Decision-making under uncertainty in dynamic settings: an experimental study," .

Rangel, A. (2008): "The computation and comparison of value in goal-directed choice," *Neuroeconomics: Decision-making and the brain. P. Glimcher, C. Camerer, E. Fehr, & R. Poldrack (eds). New York: Elsevier.*

Reutskaja, E., R. Nagel, C. Camerer, and A. Rangel (forthcoming): "Search Dynamics in Consumer Choice under Time Pressure: An Eye-Tracking Study," *American Economic Review.*

Rushworth, M., and T. Behrens (2008): "Choice, uncertainty and value in prefrontal and cingulate cortex," *Nature neuroscience*, 11(4), 389–397.

Samejima, K., K. Doya, Y. Ueda, and M. Kimura (2004): "Estimating internal variables and parameters of a learning agent by a particle filter," *Advances in Neural Information Processing Systems*, 16.

Schultz, W., P. Dayan, and P. Montague (1997): "A neural substrate of prediction and reward," *Science*, 275(5306), 1593.

Sutton, R., and A. Barto (1998): *Reinforcement Learning.* MIT Press.

Wang, J., M. Spezio, and C. Camerer (forthcoming): "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth-Telling and Deception in Sender-Receiver Game," *American Economic Review.*

Yoshida, W., and S. Ishii (2006): "Resolution of uncertainty in prefrontal cortex," *Neuron*, 50(5), 781–789.

# A Appendix: Additional details on computation of nonparametric, Bayesian, and RL learning rules

In section 5, we compared belief dynamics in the nonparametric model ($X^*$) with counterparts in other two benchmark learning models, the Bayesian belief ($B^*$) and the valuation in the reinforcement learning model ($V_b - V_g$). Here we provide additional details for how the beliefs for each of the three models were computed.

## A.1 Belief dynamics $X^*$ in the nonparametric model

The values of $X^*$, the belief process in our nonparametric learning model, were obtained by maximum likelihood. For each block, using the estimated choice and measurement probabilities, as well as the learning rules, we chose the path of beliefs $\{X_t^*\}_{t=1}^{25}$ which maximized $P(\{X_t^*\} \,|\, \{Y_t, Z_t, R_t\})$, the conditional ("posterior") probability of the beliefs, given the observed sequences of choices, eye-movements, and rewards. Because

$$P(\{X_t^*, Z_t\} \,|\, \{Y_t, R_t\}) = P(\{X_t^*\} \,|\, \{Z_t, R_t\}) \cdot P(\{Z_t\} \,|\, \{Y_t, R_t\}),$$

where the second term on the RHS of the equation above does not depend on $X_t^*$, it is equivalent to maximize $P(\{X_t^*, Z_t\} \,|\, \{Y_t, R_t\})$ with respect to $\{X_t^*\}$. Because of the Markov structure, the joint log-likelihood factors as:

$$\log L(\{X_t^*, Z_t\}|\{Y_t, R_t\}) = \sum_{t=1}^{24} \log \left[P(Z_t|X_t^*)P(X_{t+1}^*|X_t^*, R_t, Y_t)\right] + \log(P(Z_{25}|X_{25}^*)). \quad (10)$$

We plug in our nonparametric estimates of $P(Z|X^*)$ and $P(X_{t+1}^*|X_t^*, R_t, Y_t)$ into the above likelihood, and optimize it over all paths of $\{X_t^*\}_{t=1}^{25}$ with the initial condition restriction $X_1^* = 2$ (beliefs indicate "not sure" at the beginning of each block). To facilitate this optimization problem, we derive the optimal sequence of beliefs using a dynamic-programming (Viterbi) algorithm; cf. Ghahramani (2001).

In the above, we treated the choice sequence $\{Y_t\}$ as exogenous, and left the choice probabilities $P(Y_t|X_t^*)$ out of the log-likelihood function (10) above. This was because, given our estimates that $P(Y_t = 1|X_t^* = 1) \approx P(Y_t = 2|X_t^* = 3) \approx 1$ in Table 4, maximizing with respect to these choice probabilities would leads to estimates of beliefs $\{X^*\}$ which closely coincide with observed choices; we wished to avoid such an artifically good "fit" between the beliefs and observed choices.

For robustness, however, we also estimated the beliefs $\{X^*\}$ under two alternative scenarios: (i) treating the choice sequence $\{Y_t\}$ as endogenous, and hence including the choice probabilities $P(Y_t|X_t^*)$ in the likelihood function; (ii) treating both $\{Y_t, Z_t\}$ as exogenous, and hence omitting both the choice probabilities $P(Y_t|X_t^*)$ and the measurement probabilities $P(Z_t|X_t^*)$ from the likelihood function.

Not surprisingly, the correlation between choices and beliefs $\text{Corr}(Y_t, X_t^*) = 0.99$ under (i), while under (ii) the correlation falls to 0.56. However, in both of these alternative specifications, we still find that $\text{Corr}(X_t^*, V_t^*) > \text{Corr}(X_t^*, B_t^*)$ – that is, the nonparametric beliefs are "closer" to the RL model than the Bayesian model. Thus this finding appears robust across a number of different approaches to recovering the nonparametric beliefs $\{X_t^*\}$.

## A.2 Bayesian Learning Model

A Bayesian learner uses Bayes rule to update her beliefs. Let $B_t^*$ denote the belief, or prior probability, that the blue slot machine is the good one at the start of the trial $t$. After her choice, she observes reward $R_t$. Let $B'^*_t$ denote the posterior belief, the probability that the blue slot machine is the good one after $Y_t$ is chosen, and the reward $R_t$ is realized. The posterior probability is derived using Bayes rule:

$$B'^*_t = \frac{P(R_t|Y_t, S_t = 1) \cdot B_t^*}{P(R_t|Y_t, S_t = 1) \cdot B_t^* + P(R_t|Y_t, S_t = 2) \cdot (1 - B_t^*)} \tag{11}$$

At the end of each trial, the state $S_t$ may change with 15% probability. The Bayesian learner takes this into account, so that the prior probability on "blue" at the start of trial $t+1$ is the posterior probabilities $(B')_t^*$ weighted by the state transition probabilities:

$$B_{t+1}^* = P(S_{t+1} = 1|S_t = 1) \cdot B'^*_t + P(S_{t+1} = 1|S_t = 2) \cdot (1 - B'^*_t). \tag{12}$$

In this way, given the initial beliefs $B_1 = 0.5$, we can use Eqs. (11) and (12) to compute the sequence of Bayesian beliefs, $\{B_t^*\}$, corresponding to the observed sequences of choices and rewards $\{Y_t, R_t\}$. The corresponding choice rule from the Bayesian model would be to choose "blue" at trial $t$ iff $B_t^* \geq 0.5$.

## A.3 Reinforcement Learning Model

We employ a variant of the TD (Temporal-Difference)-Learning models (Sutton and Barto (1998), section 6). The value of an action is learned by the reward that is expected after taking that action. Let $V^t_{b(g)}$ denote the "current" (ie. beginning of trial $t$) action value function for the blue (green) slot machine. The value updating rule for a "One-step TD-Learning" model is defined as:

$$V^{t+1}_{Y_t} \longleftarrow V^t_{Y_t} + \alpha\delta_t. \tag{13}$$

where $Y_t$ denotes the choice taken in trial $t$, $\alpha$ denotes the learning rate, and $\delta_t$ denotes the "prediction error" for trial $t$ (defined below). For greater model flexibility, we allow the parameter $\alpha$ to take two values, following positive and negative rewards. The prediction error $\delta_t$ is equal to

$$\delta_t = (R_t + \gamma E[V^t_{Y_{t+1}}|t]) - V^t_{Y_t} \tag{14}$$

the difference between $(R_t + \gamma E[V^t_{Y_{t+1}}|t])$ (the observed reward in trial $t$ plus the discounted expected value from the next trial), and $V^t_{Y_t}$ (the current expected valuation). We assume the discount factor $\gamma = 0.9$. For instance, for $Y_t = 2$ (for "blue"), then the TD learning rule implies that $V_b$ is updated by an amount equal to the prediction error $\delta_t$, weighted by the learning parameter $\alpha$ (with larger values of $\alpha$ indicating an increased sensitivity to the outcome of trial $t$). For trial $t$, there is no updating of the valuation for the choice that was not taken.

The variant of TD-Learning (SARSA, short for "State-Action-Reward-State-Action") used here (Sutton and Barto (1998), p. 149) computes the expected value function $E[V_{Y_{t+1}}|t]$ using the current choice probabilities of choosing the future action $Y_{t+1}$ (which is unknown at trial $t$). Let $P^t_c$ denote the current probability of choosing action $c$. We adopt the conventional "softmax" (ie. logit) choice probability function with the inverse temperature parameter $\beta$:

$$P^t_c = \frac{e^{\beta V^t_c}}{\sum_{c'} e^{\beta V^t_{c'}}} \tag{15}$$

With this functional form for the choice probabilities, the expected value function from trial $t+1$ is computed as,

$$E[V_{Y_{t+1}}|t] = \sum_{c' \in (b,g)} P^t_{c'} V^t_{c'}. \tag{16}$$

To obtain estimates for $\beta$ and the two $\alpha$'s, we apply maximum likelihood estimation. The estimates we obtained from the data were:

$$\begin{aligned}
&\beta = 0.7584 \\
&\alpha \text{ for large reward } (R_t = 2) = 1.6531 \\
&\alpha \text{ for small reward } (R_t = 1) = 1.0552.
\end{aligned} \tag{17}$$

We plug in these values into Eqs. (13), (14), (15) and (16) to derive a sequence of valuations $\left\{V_t^* \equiv V_b^t - V_g^t\right\}$. The choice function (Eq. (15)) can be rewritten as a function of the difference $V_t^*$; i.e. the choice probability for the blue slot machine is,

$$P_b^t = \frac{e^{\beta(V_b^t - V_g^t)}}{1 + e^{\beta(V_b^t - V_g^t)}} = \frac{e^{\beta V_t^*}}{1 + e^{\beta V_t^*}} \tag{18}$$

and $P_g^t = 1 - P_b^t$. Hence, we see that $V_t^*$ plays a role in the TD-Learning model analogous to the belief measures $X_t^*$ and $B_t^*$ from, respectively, the nonparametric and Bayesian learning models.

# B Appendix: Some additional details on discretization

In this section, we present additional discussion on the discretization of the eye-movement measure, and some evidence that a three-valued discretization is sufficient to capture most of the variation in this measure. Let $Z_{pt}$ denote the continuous-valued eye-tracking measure, and $Z_t$ the discretized version. For the two-value discretization, we discretize as follows:

$$Z_t = \begin{cases} 1 & \text{if } Z_{pt} < 0 \\ 2 & \text{if } Z_{pt} \geq 0 \end{cases}$$

As we discussed before, since we do not find any color bias toward blue nor green, discretization of $Z_{pt}$ around 0 should be reasonable. The sample frequency for $Z_t = 1$ (green) is 2032, and that for $Z_t = 2$ (blue) is 2168 (Table 1).

For the three-value discretization, we discretize $Z_{pt}$ as follows:

$$Z_t = \begin{cases} 1 & \text{if } Z_{pt} < \text{-sid.} \\ 2 & \text{if -sid.} \leq Z_{pt} \leq \text{sid.} \\ 3 & \text{if sid.} < Z_{pt} \end{cases} \tag{19}$$

where “sid.” denotes a constant used to discretize $Z_{pt}$. The choice of the value for *sid.* does not affect the estimation results essentially. As the baseline, we set *sid.* = 0.20. However,

Table 7: Correlations ($Y$, $Z_p$) for different discretizations of eye-movement measure $Z$

"sid." is constant employed in discretization; cf. Eq. (19)

| | Size | Corr($Y_t, Z_{pt}$) |
|---|---|---|
| **Full sample** | 4200 | 0.7647 |
| **sid. = 0.20 (baseline):** | | |
| $Z_t = 1$ (green) | 1887 | 0.2845 |
| 2 (not sure) | 540 | 0.2156 |
| 3 (blue) | 1773 | 0.1706 |
| **sid.=0.05:** | | |
| $Z_t = 1$ (green) | 2015 | 0.3223 |
| 2 (not sure) | 255 | -0.0599 |
| 3 (blue) | 1930 | 0.2346 |
| **sid. = 0.40:** | | |
| $Z_t = 1$ (green) | 1725 | 0.1462 |
| 2 (not sure) | 869 | 0.2777 |
| 3 (blue) | 1606 | 0.0991 |

we do not find any difference in the estimation results either qualitatively nor significantly if we vary *sid.* from 0.05 to around 0.40, suggesting that the model is robust for different classifications. Table 7 shows the sample frequencies of the discretized measure $Z_t$ for three different values of sid.

Figure 3 contains the histogram of $Z_p$, which is noticeably trimodal, at -1, 0 and 1. Hence, discretizing $Z_{pt}$ into three values seems reasonable. Moreover, Table 7 shows the correlations between $Y$ and $Z_p$, broken up into the three ranges of $Z_p$ corresponding to the three discretized values $Z \in \{1, 2, 3\}$, and also for three different values of the *sid.* parameter. Although the correlation in the whole sample is 0.7647, the correlations within each of the three ranges of $Z_p$ are much smaller in magnitude. Because most of the variation in choices is *across* the different discretized values of $Z$, rather than within these values, it appears the three-valued discretization is sufficient.