

# Nonparametric Learning Rules from Bandit Experiments: The Eyes Have It!

September 8, 2011

## Abstract

How do people learn? We assess, in a distribution-free manner, subjects' learning and choice rules in dynamic two-armed bandit learning experiment. To aid in identification and estimation, we use auxiliary measures of subjects' beliefs, in the form of their eye-movements during the experiment. Our estimated choice probabilities and learning rules have some distinctive features; notably, subjects are more reluctant to "update down" following unsuccessful choices, than "update up" following successful choices. The profits from following the estimated learning and decision rules are smaller (by about 25% of typical experimental earnings) than what would be obtained from an optimal Bayesian learning model, but comparable to the profits from alternative non-Bayesian learning models, including reinforcement learning and a simple "win-stay" choice heuristic.

**Keywords:** learning, experiments, eye-tracking, Bayesian vs. non-Bayesian learning, nonparametric estimation

**JEL codes:** D83, C91, C14

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 experiment, where subjects must repeatedly 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. The nonparametric estimation of learning models is a new endeavor in both the experimental learning literature, as well as the empirical literature in economics and marketing in which dynamic learning models are estimated structurally using field data. Estimating the learning rules in a model-free manner allows us to assess the optimality of subjects' choices in learning experiments in a manner quite distinct from that taken in the existing literature.

A sizable literature has developed around structural estimation of learning-based models of dynamic choice. Some representative papers include R. Miller (1984), T. Erdem & M. Keane (1996), D. Akerberg (2003), G. Crawford & M. Shum (2005), T. Chan & B. Hamilton (2006), and P. Marcoul & Q. 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, there is also the opportunity to observe "auxiliary" measures of subjects' beliefs (or valuations), such as eye movements (as in K. Armel & A. Rangel (2008), and the present paper), brain activity (cf. W. Yoshida & S. Ishii (2006), E. Boorman, T. Behrens, M. Woolrich & M. Rushworth (2009) in the recent fMRI neuroscience literature), or mouse-tracking (cf. I. Brocas, J. Carrillo, S. Wang & C. Camerer (2009)).

Because of this additional 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, non-Bayesian alternatives. An incomplete

list of papers which consider these questions includes D. Grether (1992), M. El-Gamal & D. Grether (1995), G. Charness & D. Levin (2005), C. Kuhnen & B. Knutson (2008), and E. Payzan-LeNestour & P. Bossaerts (2011). Recently, non-Bayesian *reinforcement learning* (R. Sutton & A. Barto (1998)) models have also been used to explain some observed anomalies in savings and investment behavior (eg. J. Choi, D. Laibson, B. Madrian & A. Metrick (2009), T. Odean, M. Strahilevitz & B. Barber (2004)).<sup>1</sup>

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 choice and eye-tracking 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. Moreover, we also estimate subjects' decision rules nonparametrically jointly from the observed choice and eye movement data. To our knowledge, the present paper is the first which undertakes the nonparametric estimation of structural decision models using experimental data.

Our main results are: (i) subjects' reduced-form decision rules resemble an asymmetric “win-stay/lose-randomize” rule of thumb, whereby subjects replay successful strategies, but randomize after unsuccessful ones; (ii) correspondingly, our estimates of the learning rules show that subjects are more reluctant to “update down” following unsuccessful choices, than “update up” following successful choices; such asymmetries are suboptimal relative to the rational Bayesian benchmark, and (iii) we find that that subjects' profits are, at the median, \$4 (or about two cents per choice) lower than under the Bayesian benchmark, which represents about 25% of typical experimental earnings (not including the fixed show-up fee). However, subjects' profits under the estimated choice and learning rules are comparable to the profits from alternative non-Bayesian learning models, including reinforcement learning.

---

<sup>1</sup>In the computational IO literature, such learning algorithms have also been used to ease the computational burden associated with dynamic equilibrium models, cf. A. Pakes & P. McGuire (2001), S. Imai, N. Jain & A. Ching (2009).

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, A. Hampton, P. Bossaerts & J. 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 brain scans.<sup>2</sup> Instead, our approach represents a novel application of econometric tools recently developed for the estimation of nonclassical measurement error models and dynamic discrete-choice models (Y. Hu (2008), Y. Hu & M. Shum (2008)). Essentially, we 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. We obtain a simple estimator for the learning model which involves only elementary calculations involving matrices which can be formed from the observed data.<sup>3</sup>

In the next section, 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 2, we present an econometric model of subjects’ choices in the bandit model, and discuss nonparametric identification and estimation. In Section 3, we describe the experimental data, and present our nonparametric estimates of subjects decision rules and learning rules. Section 4 contains a comparison of our estimated learning rules to “standard” learning rules, including those from the Bayesian and non-Bayesian reinforcement-learning models. Section 5 concludes.

---

<sup>2</sup>Other papers utilizing a similar methodological framework include T. Behrens, M. Woolrich, M. Walton & M. Rushworth (2007), Boorman et al. (2009), N. Daw, J. O’Doherty, P. Dayan, B. Seymour & R. Dolan (2006), Yoshida & Ishii (2006).

<sup>3</sup>More broadly, because subjects’ underlying beliefs are unobserved and also serially correlated over time, learning models are a particular case of nonlinear “hidden state Markov” models, which are typically quite challenging to estimate (cf. Z. Ghahramani (2001) and P. Arcidiacono & R. Miller (2006)). Relatedly, K. Samejima, K. Doya, Y. Ueda & M. Kimura (2004) consider Bayesian estimation of a reinforcement learning model using sequential Monte Carlo (“particle filtering”) methods.

# 1 Two-armed bandit “reversal learning” experiment

Our experiments are adapted from the “reversal learning” experiment used in Hampton, Bossaerts & O’Doherty (2006). In the experiments, subjects make repeated choices between two actions (which we call interchangeably “arms” or “slot machines” in what follows): in trial  $t$ , the subject chooses  $Y_t \in \{1(= \text{“green”}), 2(= \text{“blue”})\}$ . The rewards generated by these two arms are changing across trials, as described by the state variable  $S_t \in \{1, 2\}$ , which 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 reward process is

$$R_t = \begin{cases} \pm\$0.50 \text{ with prob. } 50\% \pm 20\% & \text{if good arm chosen} \\ \pm\$0.50 \text{ with prob. } 50\% \mp 10\% & \text{if bad arm chosen.} \end{cases} \quad (1)$$

For convenience, we use the notation  $R_t = 1$  to denote the negative reward (-\$0.50), and  $R_t = 2$  to denote the positive reward (\$0.50).

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<sup>4</sup>

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

(2)

Because  $S_t$  is not observed by subjects, and is serially-correlated over time, subject have an opportunity to learn and update their beliefs about the current state on the basis of past

---

<sup>4</sup>This aspect of our model differs from Hampton, Bossaerts & O’Doherty (2006), who make the non-Markovian assumption that the state  $S_t$  changes with probability 25% after a subject has chosen the good arm four successive times. Estimating such non-Markovian models would, typically, require including another state variable, which describe how uncertain the subject is at any point in time about the underlying state (such as the variance of rewards). In principle, our estimation method can be extended to allow for these additional state variables, but since we take a nonparametric approach, a much larger sample size (far beyond typical samples sizes in experimental work) would be required to obtain reasonable estimates.

rewards. Moreover, because  $S_t$  changes randomly over time, so that the identity of the good arm varies across trials, this is called a “probabilistic reversal learning” experiment.

## 1.1 Optimal decision-making in reversal learning model

Before proceeding to describe the experimental data, we consider how subjects should optimally make decisions in the dynamic reversal-learning model used in our experiments. The qualitative features of the optimal decision and belief updating rules presented here will motivate the assumptions which underlie the empirical learning model which we estimate in this paper. As in the experiments, we consider a finite (25 period) dynamic optimization problem, in which each subject aims to choose a sequence of actions to maximize expected rewards  $\mathbb{E} [\sum_{t=1}^{25} R_t]$ . (The details of this model are given in Appendix A.)

Let  $B_t^*$  denote the probability (given by Bayes’ Rule) denote the probability that a subject places on “green” being the good arm in period  $t$ , conditional on the whole experimental history up to then. We evaluate the optimal decision rules – the mapping from period  $t$  beliefs  $B_t^*$  to a period  $t$  choice – in this dynamic Bayesian learning model by computer simulation. Importantly, we accommodate nonstationarity in the problem, in that our simulations allow the decision rules to differ arbitrarily across periods. This permits the relationship between subjects’ choices and their beliefs  $B_t^*$  to vary across periods, depending perhaps on the periods remaining in the experiment, or to allow for history dependence in either choices or the belief-updating rule. An important maintained assumption in this paper is that subjects’ decision rules are solely a function of the current state probabilities  $B_t^*$ , so that by allowing the decision rules to vary across periods in these simulations, we can assess the restrictiveness of such an assumption.

The important qualitative features of optimal decision-making are summarized in the optimal decision-rules, which we plot in Figure 2 for four periods  $t = 1, 10, 20, 25$ . Two features are apparent. First, we see that the decision rules are identical across all the periods, indicating that they are *stationary*. Second, the optimal decision rule takes a simple form: in each period, the subject chooses simply blue once the current belief that the blue arm is “good”

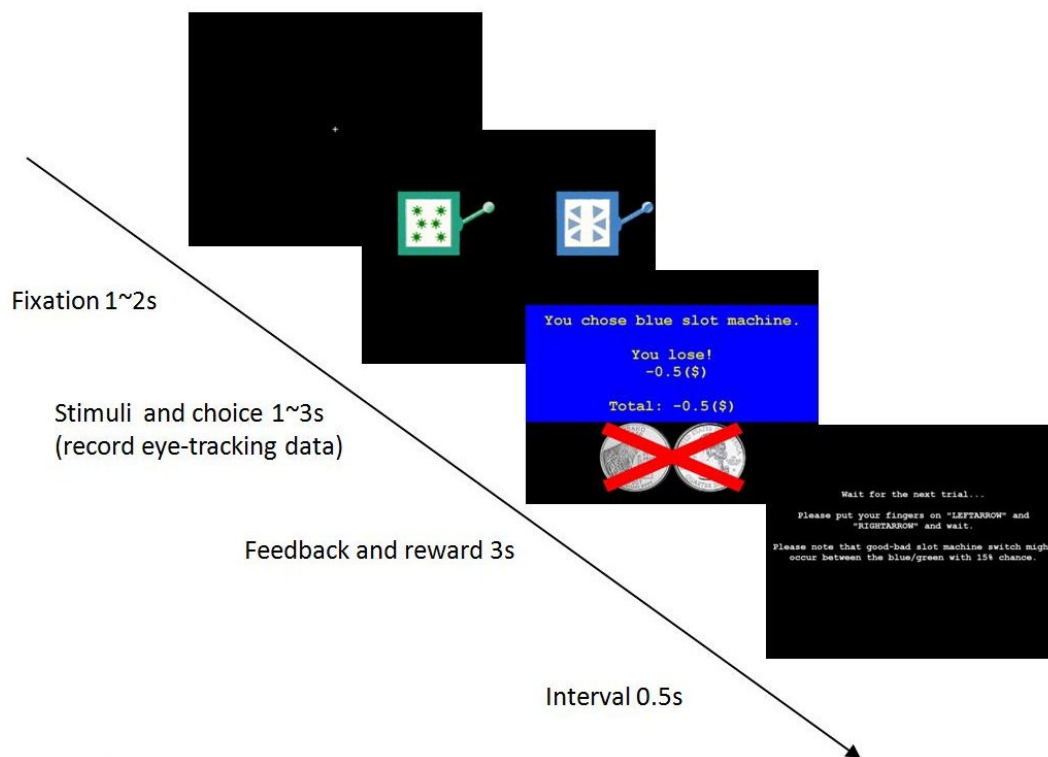
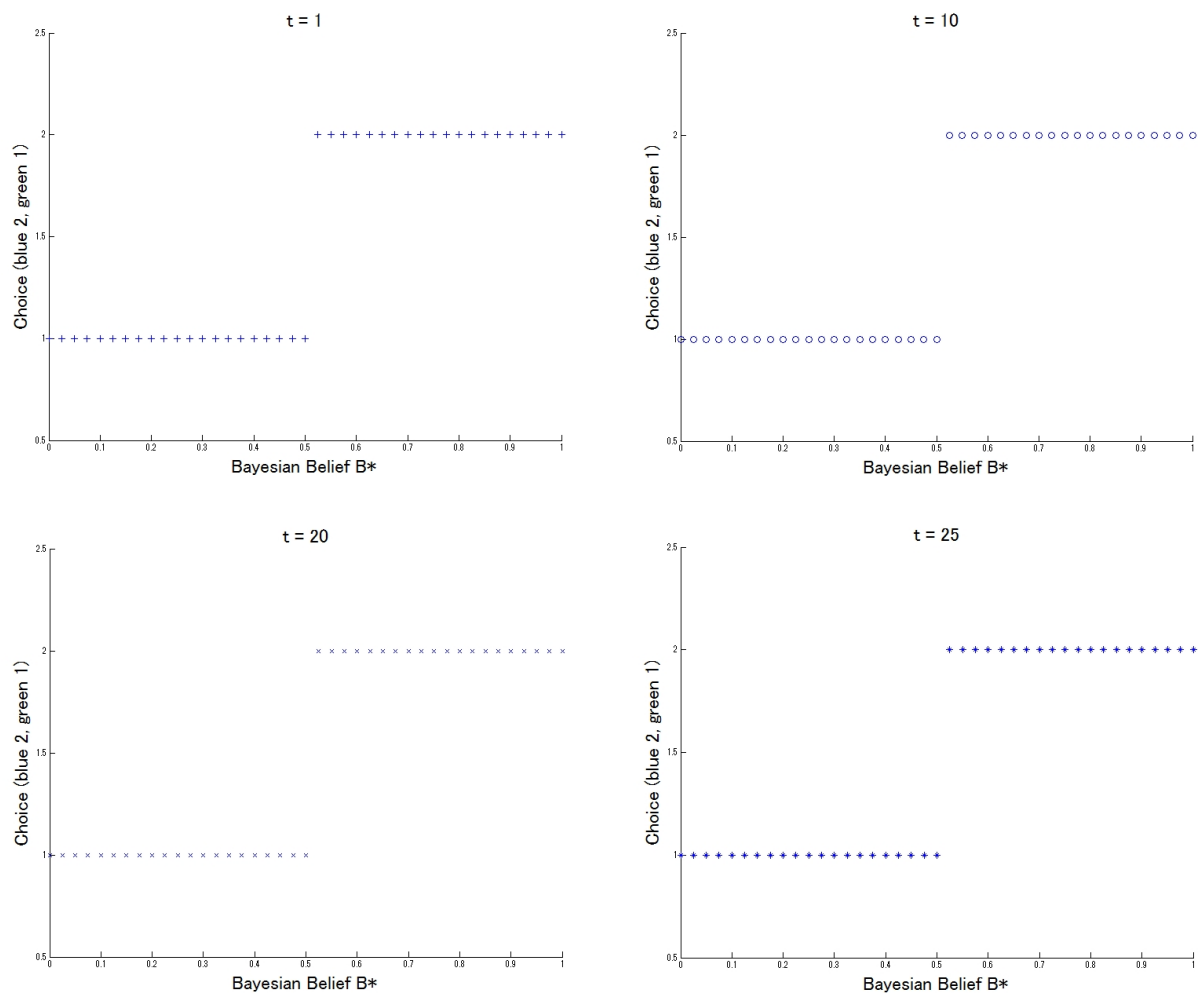


Figure 1: Timeline of a trial

After a fixation on the cross (top screen), two slot machines are presented (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 a slot machine may switch from "good" to "bad" (and vice versa) with probability 15%.

Figure 2: Optimal decision rules in reversal learning model

Plotted for periods  $t = 1, 10, 20, 25$ .

exceeds 50%. This is a *myopic* decision rule.

Both of these features – stationarity and myopia of decision rules – are specific to the reversal-learning setup considered here, and differs in important ways from optimal decision-making in the standard multi-armed bandit (MAB) problem (cf. J. Gittins & G. Jones (1974), J. Banks & R. 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 in the standard MAB model features exploration (or “experimentation”), which recommends sacrificing current rewards to achieve longer-term payoffs; this makes simple myopic decision-making (choosing the bandit which currently has the higher expected reward) suboptimal. In the reversal learning setting, however, the states of the bandits are negatively related, so that positive information about one slot machine implies negative information about the other. Apparently, as shown by these optimal decision rules, this eliminates most of the incentives for subjects to experiment.

Moreover, in a finite-horizon decision environment, such as the experiments considered here, the value of information decreases exogenously as the final period approaches, resulting in reduced incentives for experimentation; this implies a nonstationary decision rule. Under reversal learning, however, the lack of experimentation leads to stationary decision rules, even in a finite horizon problem.

## 1.2 Experimental data: preliminary analysis

The experiments were run over several weeks 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,<sup>5</sup> each playing for 200 rounds (broken up into 8 blocks of 25 trials). Most of the subjects completed the experiment within 40 minutes, including instruction and practice

---

<sup>5</sup>Community members consisted of spouses of students at either Caltech or Pasadena City College (a two-year junior college). 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 noticeable differences in the results across these classes of subjects.

sessions. Subjects were paid a fixed show-up fee (\$20), in addition to the amount won during the experiment, which was \$14.20 on average.<sup>6</sup>

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 occurring in each trial. 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  $\tilde{Z}_t$ , 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.<sup>7</sup>

For each subject, and each round  $t$ , we observe the data  $(Y_t, S_t, R_t, Z_t)$ . Table 1 presents some summary statistics of the data. The top panel shows that, across all subjects and all trials, “green” (2108 choices) and “blue” (2092 choices) are chosen in almost-equal proportions. Moreover, from the second panel, we see that subjects obtain the high reward 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.<sup>8</sup> Hence, subjects appear to be “trying”, which motivates our analysis of their learning rules. On the other hand, simulation of the optimal Bayesian decision rules (discussed above) show that the success rate from using the optimal decision rule is only 58.4%, which is just slightly higher than the in-sample success rate found in the experiments. It appears, then, that in the reversal learning setting, the success rate intrinsically varies quite narrowly between 55% and 58.4%.

In Table 2, we present the conditional probabilities of choices in period  $t$ , conditional on choices and rewards from the previous period  $(Y_t|Y_{t-1}, R_{t-1})$ . This can be interpreted as

---

<sup>6</sup>For comparison, purely random choices would have earned \$10 on average.

<sup>7</sup>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).

<sup>8</sup>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.

Table 1: Summary statistics for experimental data

	1( <span style="color: green;">green</span> )	2( <span style="color: blue;">blue</span> )
$Y$ : subjects' choices	2108	2092

	1 (\$0.50)	2 (-\$0.50)
$R$ : rewards	2398	1802

	mean	median	upper 5%	lower 5%
$\tilde{Z}$ : eye movement measure <sup>a</sup>	-0.0309	0	1.3987	-1.4091
$RT$ : reaction time ( $10^{-2}$ secs)	88.22	59.3	212.2	36.8

---

<sup>a</sup>Defined in Eq. (3)

a “reduced-form” decision rule for the subjects. The top row in that table contains the reduced-form probabilities of choosing the green arm. Looking at the second (fourth) entry in this row, we see that after a successful choice of green (blue), a subject replays this strategy with probability 0.86 (0.88=1-0.12). Thus subjects appear to replay successful strategies, corresponding to a “win-stay” rule-of-thumb.

However subjects appear reluctant to give up *unsuccessful* strategies. The probability of replaying a strategy after an unsuccessful choice of the same strategy is around 50% for both the blue and green choices (ie. the first and third entries in this row). Thus, subjects tend to randomize after unsuccessful strategies. As far as we are aware, such an “asymmetric” choice rule is new in the literature; moreover, as we will see below, this is echoed in the “asymmetric” belief-updating rule which we estimate.

In the remainder of Table 2, we also present the same choice probabilities, calculated for each subject individually. There is some degree of heterogeneity in subjects’ strategies. Looking at columns 2 and 4 of the table, we see that, for the most part, subjects pursue a “win-stay” strategy: the probabilities in the second column are mainly  $\gg 50\%$ , and those in the fourth column are most  $\ll 50\%$ . However, looking at columns 1 and 3, we see that there is significant heterogeneity in subjects’ choices following a low reward. In

these cases, randomization (which we classify as a choice probability between 40-60%) is the modal strategy among subjects; strikingly, however, a number of subjects continue replaying an unsuccessful strategy: for examples, subjects 3,8, and 11 continue to choose “green” with probabilities of 79%, 89% and 79% even after a previous choice of green yielded a negative reward.<sup>9</sup>

One common feature of the choice strategies across all subjects is that choices are serially correlated across periods, conditional on rewards. This serial correlation is very informative for identifying the beliefs  $X_t^*$ . Essentially, we present a model below in which serial correlation in choices across periods arises due to beliefs — thus, beliefs (which are unobserved to the researcher) are the reason for serial correlation of choices. We will discuss this in more detail in the next section, in which the empirical model is presented formally.

### 1.3 Remarks on eye-tracking measure

Eye-tracking is still a relatively novel tool in economics; recently, it has been employed to assess subjects thinking processes in various decision environments: to determine how subjects detect truth-telling or deception in sender-receiver games (J.T. Wang, M. Spezio & C.F. Camerer (2010)); how consumers evaluate comparatively a huge number of commodities, as in a supermarket setting (E. Reutskaja, R. Nagel, C.F. Camerer & A. Rangel (2011)); and the relationship between visual attention (as measured by eye-fixations) and valuation of commodities in choice tasks (cf. I. Krajbich, C. Armel & A. Rangel (2010), Armel & Rangel (2008), K. Armel, A. Beaumel & A. Rangel (2008), A. Rangel (2008)).<sup>10</sup>

Our use of eye movements in this paper is predicated on an assumption that gaze is related to beliefs of expected rewards. This is motivated by some recent results in behavioral neuroscience. S. Shimojo, C. Simion, E. Shimojo & C. Scheier (2003) studied this in binary “face

---

<sup>9</sup>In the reversal learning model, however, such a strategy is not obviously irrational; because the identity of the good arm changes exogenously across periods, an arm that was bad last period (ie. yielding a low reward) may indeed be good in the next period.

<sup>10</sup>Eye-tracking has also been used in marketing studies to evaluate the relationship between visual attention to advertisements and subsequent sales of advertised items (eg. J. Zhang, M. Wedel & R. Pieters (2009)).

Table 2: “Reduced-form” decision rule:  $P(Y_t = 1(\text{green})|Y_{t-1}, R_{t-1})$ 

$(Y_{t-1}, R_{t-1})$ :	(1,1)	(1,2)	(2,1)	(2,2)
All Subjects:	0.5075 (0.0169)	0.8652 (0.0094)	0.5089 (0.1169)	0.1189 (0.0090)
Subject1:	0.1799 (0.0655)	0.5192 (0.0684)	0.8128 (0.0595)	0.364 (0.0603)
Subject2:	0.1051 (0.0498)	0.9820 (0.0171)	0.9449 (0.0381)	0 (0)
Subject3:	0.7938 (0.0591)	0.9859 (0.0136)	0.3340 (0.0871)	0 (0)
Subject4:	0.3244 (0.0704)	0.8796 (0.0514)	0.6492 (0.0726)	0.0610 (0.0283)
Subject5:	0.0419 (0.0292)	0.8796 (0.0236)	0.6492 (0.0325)	0.0610 (0.0461)
Subject6:	0.2570 (0.0652)	0.7498 (0.0592)	0.8159 (0.0602)	0.2021 (0.0532)
Subject7:	0.5792 (0.0751)	0.9242 (0.0371)	0.4647 (0.0731)	0.0796 (0.0379)
Subject8:	0.8931 (0.0496)	0.9803 (0.0186)	0.1013 (0.0482)	0.0165 (0.0163)
Subject9:	0.6377 (0.0831)	1.0000 (0)	0.2741 (0.0655)	0 (0)
Subject10:	0.1986 (0.0622)	0.9344 (0.0352)	0.8037 (0.0587)	0 (0)
Subject11:	0.7859 (0.0575)	1.0000 (0)	0.4306 (0.0870)	0 (0)
Subject12:	0.5883 (0.0841)	0.9262 (0.0406)	0.3741 (0.0733)	0.0131 (0.0129)
Subject13:	0.6741 (0.0705)	0.8907 (0.0462)	0.1962 (0.0581)	0.2085 (0.0539)
Subject14:	0.4730 (0.0831)	0.6147 (0.0653)	0.5363 (0.0735)	0.3842 (0.0664)
Subject15:	0.6759 (0.0761)	0.9789 (0.0206)	0.3351 (0.0714)	0 (0)
Subject16:	0.4595 (0.0715)	0.9135 (0.0316)	0.5443 (0.0742)	0.1953 (0.0666)
Subject17:	0.6358 (0.0660)	0.5202 (0.0706)	0.5322 (0.0780)	0.4644 (0.0748)
Subject18:	0.6333 (0.0834)	1.0000 (0)	0.2901 (0.0734)	0 (0)
Subject19:	0.6144 (0.0702)	0.8197 (0.0444)	0.5808 (0.0806)	0.2013 (0.0625)
Subject20:	0.3699 (0.0858)	0.5741 (0.0707)	0.3699 (0.0665)	0.3554 (0.0621)
Subject21:	0.6990 (0.0658)	0.9602 (0.0274)	0.2934 (0.0693)	0.0177 (0.0171)

Note: standard errors (in parentheses) computed using 1000 bootstrap resamples

choice” tasks, in which subjects are asked to choose one of the two presented faces on the basis of various criteria. (Our two-armed bandit task is very similar in construction.) These authors find that, when subjects are asked to choose a face based on attractiveness, their eye movements are directed to the preferred face and remained there longer. Interestingly, the relationship between gaze duration and the chosen face becomes significantly weaker when subjects are asked to choose a face based on shape and “unattractiveness”. This strongly suggests that directed gaze duration reflects preferences, rather than choices.

This work echoes primate experiments reported in J. Lauwereyns, K. Watanabe, B. Coe & O. Hikosaka (2002) and R. Kawagoe, Y. Takikawa & O. Hikosaka (1998) (see the survey in O. Hikosaka, K. Nakamura & H. Nakahara (2006)), which shows that primates tend to direct their gaze at locations where rewards are available. They also establish a physiological basis for this relationship, by showing a connection between eye movements and reward-sensitive neuronal activities in the basal ganglia part of the brain. These results, which link gaze direction and duration with expected rewards, provide some precedence and justification to our use of eye movements as “noisy measures” of beliefs, which likewise reflect perceptions of expected rewards from the slot machines.<sup>11</sup>

Based on the papers above, we define  $\tilde{Z}_{it}$ , our raw eye-movement measure, as the difference in the gaze duration directed at the blue and green slot machines, normalized by the total reaction time:

$$\tilde{Z}_t = (Z_{b,t} - Z_{g,t})/RT_t; \quad (3)$$

that is, for trial  $t$ ,  $Z_{b(g),t}$  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.<sup>12</sup> Thus,  $\tilde{Z}_t$  measures how much longer a subject looks at the blue slot machine than

---

<sup>11</sup>An alternative to using eye movements to proxy for beliefs would have been to elicit beliefs (as in Y. Nyarko & A. Schotter (2002)). However, given the length of our experiments (8 trials of 25 periods each), and our need to have beliefs for each period, it seemed infeasible to elicit beliefs. Indeed, in our pilot experiments, we tried eliciting beliefs randomly after some periods, and found that this made the experiments unduly long.

<sup>12</sup>Furthermore, in order to control for subject-specific heterogeneity, we normalize  $\tilde{Z}_t$  across subjects by dividing by the subject-specific standard deviation of  $\tilde{Z}_t$ , across all rounds for each subject.

the green one during the  $t$ -th trial, with a larger (smaller) value of  $\tilde{Z}_t$  implying longer fixation time at the blue (green) slot machine. Summary statistics on this measure are given in the bottom panel of Table 1. There, we see that the average reaction time is 0.88 seconds, and that the median value of  $\tilde{Z}_t$  is zero, implying an equal amount of time directed to each of the two slot machine.<sup>13</sup>

Figure 5 in the Appendix contains the scatter plot of  $Z_{b,t}$  versus  $Z_{g,t}$ . In our empirical work, we will discretize the eye-movement measure  $\tilde{Z}_t$ ; to avoid confusion, in the following we use  $\tilde{\tilde{Z}}_t$  to denote the undiscretized eye-movement measure, and  $Z_t$  the discretized measure, which we describe below.

## 2 Empirical Model

In this section, we describe our econometric model of dynamic decision-making in the two-armed bandit experiment described above, and also discuss the identification and estimation of this model. Importantly, most of the crucial assumptions of the model are motivated by the structure of the optimal decision rules and learning (belief updating) rules, as described in Section 1.1. That is, we do not consider the whole gamut of learning models here, but restrict attention to models which are “close” to optimal in that the structure of the learning and decision rules are the same as in the optimal model; however, the rules themselves are allowed to be different.

---

<sup>13</sup>Following the suggestions of a referee, we also considered an alternative definition of the eye movement measure  $\tilde{\tilde{Z}}_t = (Z_{b,t} - Z_{g,t}) / (Z_{b,t} + Z_{g,t})$ , in which the time spent gazing at the middle of the screen (which is  $RT_t - Z_{b,t} - Z_{g,t}$ ) is not included in the denominator. This allows for the possibility that the time spent gazing in the middle may be indicative of “contemplation”, and may lead to stronger subsequent beliefs. We found that the estimation results for the choice probabilities and learning rules from this alternative specification (which are available from the authors upon request) are quite similar to the results from our standard specification, which are reported below. This suggests that the time spent gazing in the middle of the screen is not that informative about the evolution of subjects’ beliefs. Similarly, another referee suggested that the absolute reaction time  $RT_t$  itself could be included in the definition of eye movements. However, we found that the absolute value of  $\tilde{\tilde{Z}}_t$  is inversely related to  $RT_t$ ; this suggests that our measure  $\tilde{\tilde{Z}}_t$  appears to capture or contain the information in reaction time.

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.<sup>14</sup> 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  $\{1, 2, \dots, K\}$ . 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  $P(Y_t = y|X_t^*)$  varies across different values of  $X_t^*$  (ie. beliefs affect actions).*

Because we interpret the unobserved variables  $X_t^*$  here as a measurement of subjects’ *current* beliefs regarding which arm is currently the “good” one, the choice probability  $P(Y_t|X_t^*)$  can be interpreted as that which arises from a “myopic” choice rule. As we remarked before, in Section 1.1, such an interpretation is justified by the simulation of the optimal choice rules under the reversal learning setting, which showed that these rules are myopic and depend only on current beliefs.

Furthermore, Assumption 1 embodies an important exclusion restriction that, conditional on beliefs  $X_t^*$ , the observed action  $Y_t$  is independent of the eye movement  $Z_t$ . As we will see below, this is a critical identification assumption which pins down the beliefs  $X_t^*$  in the empirical model.

**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  $P(X_t^*|X_{t-1}^*, R_{t-1}, Y_{t-1})$ .*

This assumption is motivated by the structure of the optimal Bayesian belief-updating rule (cf. Eq. (10) in Appendix A), in which the period  $t$  beliefs depend only on the past beliefs,

---

<sup>14</sup> $X_t^*$  corresponds to the prior beliefs  $p_t$  from the previous section except that, further below, we will discretize  $X_t^*$  and assume that it is integer-valued. Therefore, to prevent any confusion, we will use distinct notation  $p_t$ ,  $X_t^*$  to denote, respectively, the beliefs in the theoretical vs. the empirical model.

actions, and rewards in period  $t - 1$ . However, we allow the exact form of the learning rule to deviate from the exact Bayes formula.

**Assumption 3** *The eye movement measure  $Z_t$  is a noisy measure of beliefs  $X_t^*$ :*

(i) *Eye movements are serially uncorrelated conditional on beliefs:  $P(Z_t|X_t^*, Y_t, Z_{t-1}) = P(Z_t|X_t^*)$ .*

(ii) *For all  $t$ , the  $K \times K$  matrix  $\mathbf{G}_{Z_t|Z_{t-1}}$ , with  $(i, j)$ -th entry equal to  $\Pr(Z_t = i|Z_{t-1} = j)$ , is invertible.*

(iii)  *$E[Z_t|X_t^*]$  is increasing in  $X_t^*$ .*

As with Assumption 1, this assumption involves an important exclusion restriction that, conditional on  $X_t^*$ , the eye movement  $Z_t$  in period  $t$  is independent of  $Z_{t-1}$ . This serial independence assumption is, to some extent, imposed by construction in the experimental setup, because we require subjects to “fix” their gaze in the middle of the computer screen at the beginning of each period. This should remove any inherent serial correlation in eye movements which is not related to the learning task.<sup>15</sup>

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  $Z_t|Z_{t-1}$ ; hence it is testable. 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.

**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 and trials  $t$ .*

This “stationarity” assumption justifies pooling the data across all subjects and trials for estimating the model. As with the other assumptions, it is motivated by the structure of

---

<sup>15</sup>At the same time, we have also estimated models in which we allow  $Z_t$  and  $Z_{t-1}$  to be correlated, even conditional on  $X_t^*$ . These are reported in Appendix D. The results there show that the results are quite similar, for different values of  $Z_{t-1}$ , which imply that Assumption 3 is quite reasonable.

optimal decision-making discussed in Section 1.1 above, where both the Bayesian belief-updating rule (Eq. (10) in Appendix A) and optimal choice rules in Figure 2 are indeed stationary.

## 2.1 Identification

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

$$f(Y_t, Z_t, X_t^* | Y_{<t}, Z_{<t}, R_{<t}, X_{<t}^*) = f(Y_t, Z_t, X_t^* | Y_{t-1}, R_{t-1}, X_{t-1}^*). \quad (4)$$

Abusing terminology somewhat, we call this a “first-order Markov” property, because the model exhibits only a one-period history dependence:

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

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

The unknown functions we want to identify and estimate are:

- (i)  $f(Y_t | X_t^*)$ , the *choice probabilities*;
- (ii) the *learning rule*  $f(X_t^* | X_{t-1}^*, Y_{t-1}, R_{t-1})$ ; and
- (iii) the *measurement probabilities*  $f(Z_t | X_t^*)$ , the mapping between the auxiliary measure  $Z_t$  and the unobserved beliefs  $X_t^*$ .

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  $\mathbf{P}(\mathbf{Y}_t|\mathbf{X}_t^*)$  and measurement probabilities  $\mathbf{P}(\mathbf{Z}_t|\mathbf{X}_t^*)$ .** Consider the joint density  $f(Z_t, Y_t|Z_{t-1})$ , which is solely a function of variables observed in the data. We can factor this density as follows:

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

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:

$$\begin{aligned}
\mathbf{A}_{y, Z_t|Z_{t-1}} &= [f_{Y_t, Z_t|Z_{t-1}}(y, i|j)]_{i,j} \\
\mathbf{B}_{Z_t|X_t^*} &= [f_{Z_t|X_t^*}(i|k)]_{i,k} \\
\mathbf{C}_{X_t^*|Z_{t-1}} &= [f_{X_t^*|Z_{t-1}}(k|j)]_{k,j} \\
\mathbf{D}_{y|X_t^*} &= \begin{bmatrix} f_{Y_t|X_t^*}(y|1) & 0 & 0 \\ 0 & f_{Y_t|X_t^*}(y|2) & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & f_{Y_t|X_t^*}(y|K) \end{bmatrix} \tag{5}
\end{aligned}$$

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

$$\mathbf{G}_{Z_t|Z_{t-1}} = [f_{Z_t|Z_{t-1}}(i|j)]_{i,j}. \tag{6}$$

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

$$\mathbf{A}_{y,Z_t|Z_{t-1}} \mathbf{G}_{Z_t|Z_{t-1}}^{-1} = \mathbf{B}_{Z_t|X_t^*} \mathbf{D}_{y|X_t^*} \mathbf{B}_{Z_t|X_t^*}^{-1}. \quad (7)$$

This is an eigenvalue decomposition of the matrix  $\mathbf{A}_{y,Z_t|Z_{t-1}} \mathbf{G}_{Z_t|Z_{t-1}}^{-1}$ , which can be computed from the observed data sequence  $\{Y_t, Z_t\}$ .<sup>16</sup> 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. (7) 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. (7) 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  $\mathbf{P}(X_{t+1}^*|X_t^*, \mathbf{R}_t, \mathbf{Y}_t)$ .** Again, start with a factorization

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

---

<sup>16</sup>Note that, from Eq. (6), the invertibility of  $\mathbf{G}$  (which is Assumption 3(i)) implies the invertibility of  $\mathbf{B}$ .

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

$$\mathbf{H}_{Z_{t+1}, y, r, Z_t} = \mathbf{B}_{Z_{t+1}|X_{t+1}^*} \mathbf{L}_{X_{t+1}^*, X_t^*, y, r} \mathbf{B}'_{Z_t|X_t^*}.$$

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

$$\begin{aligned} \mathbf{H}_{Z_{t+1}, y, r, Z_t} &= [f_{Z_{t+1}, Y_t, R_t, Z_t}(i, y, r, j)]_{i, j} \\ \mathbf{L}_{X_{t+1}^*, X_t^*, y, r} &= [f_{X_{t+1}^*, X_t^*, Y_t, R_t}(i, j, y, r)]_{i, j}. \end{aligned} \quad (8)$$

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

$$\mathbf{L}_{X_{t+1}^*, X_t^*, y, r} = \mathbf{B}_{Z_{t+1}|X_{t+1}^*}^{-1} \mathbf{H}_{Z_{t+1}, y, r, Z_t} [\mathbf{B}'_{Z_t|X_t^*}]^{-1}. \quad (9)$$

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 Estimation

Our estimation procedure mimics the two-step identification argument from the previous section. That is, for fixed values of  $(y, r)$ , we first form the matrices  $\mathbf{A}$ ,  $\mathbf{G}$ , and  $\mathbf{H}$  (as defined previously) from the observed data, using sample frequencies to estimate the corresponding probabilities. Then we obtain the matrices  $\mathbf{B}$ ,  $\mathbf{D}$ , and  $\mathbf{L}$  using the matrix manipulations in Eqs. (7) and (9).

To implement this, we assume that the eye movement measures  $Z_t$  and the unobserved beliefs  $X_t^*$  are discrete, and take three values.<sup>17</sup> One technical feature is that, because all the elements in the matrices of interest  $\mathbf{B}$ ,  $\mathbf{D}$ , and  $\mathbf{L}$  correspond to probabilities, they must take values within the unit interval. However, in the actual estimation, we found that occasionally

---

<sup>17</sup>Since the eye-movement measure  $\tilde{Z}_t$  is continuous, we must discretize it for estimation. We leave the details of our discretization procedure (including a discussion of whether three points of discretization is appropriate) in Appendix B.

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. (7) and (9), 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.<sup>18</sup>

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  $\tau = 2, \dots, 25$ . 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  $\mathbf{A}$ ,  $\mathbf{G}$ , and  $\mathbf{H}$  converge towards population counterparts.<sup>19</sup>

### 3.1 Estimation results

Tables 3 and 4 present estimation results. Both  $X_t^*$  and  $Z_t$  are discretized to take values  $\{1, 2, 3\}$ . 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”.<sup>20</sup>

---

<sup>18</sup>In principle, because we don’t impose *a priori* that the estimated probabilities must lie in  $[0, 1]$  in estimated, we could use these overidentifying restrictions to test the model. While this works as an informal “eyeball” test, developing the formal sampling theory behind such a test seems difficult, due to the complexities in characterizing the behavior of a test statistic at the boundary values of 0 and 1, and so we do not pursue it here.

<sup>19</sup>Results from Monte Carlo simulations (available from the authors on request) 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.

<sup>20</sup>We have tried to re-estimate the model allowing for more belief states ( $\geq 4$ ), but the results we obtained was not encouraging. This is due to our relatively small sample size; since our estimation approach is nonparametric, it is difficult to obtain reliable estimates with modest sample sizes. At the same time, as we pointed out above, statistical evidence indicates that it is sufficient to discretize the eye movement measure  $Z_t$  into three values, which implies that beliefs  $X_t^*$  should not take more than 3 values.

Table 3 contains the estimates of the choice and measurement probabilities.<sup>21</sup> 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 probability of choosing a color contrary to beliefs – which is called the “exploration probability” in the literature – is small, being equal to 1.3% when  $X_t^* = 1$ , and only 0.64% when  $X_t^* = 3$ .

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 split their gaze more evenly between the two colors (ie.  $Z_t = 2$ ) around 63% of the time.

The learning rule estimates are presented in Table 4. The left columns show how beliefs are updated when “exploitative” choices (ie. choices made in accordance with beliefs) are taken, and illustrate an important asymmetry in subjects’ belief-updating rules. When current beliefs indicate “green” ( $X_1^* = 1$ ) and green is chosen ( $Y_t = 1$ ), beliefs evolve asymmetrically depending on the reward: if  $R_t = 2$  (high reward), then beliefs update towards green with probability 89%; however, if  $R_t = 1$  (low reward), then belief still stay at green with probability 57%. This tendency of subjects to update up after successes, but not update down after failures also holds after a choice of “blue” (as shown in the left-hand columns of the bottom two panels in Table 4): there, subjects update their belief on blue up to 88% following a success ( $R_t = 2$ ), but still give the event blue a probability of 53% following a failure ( $R_t = 1$ ). This muted updating following failures is a distinctive feature of our learning rule estimates and, as we will see below, is at odds with optimal Bayesian belief-updating.

The results in the right-most columns describe belief updating following “explorative” (contrarian to current beliefs) choices. For instance, considering the top two panels, when current

---

<sup>21</sup>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 3: Estimates of choice and measurement probabilities

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

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

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

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%). However, the standard errors (computed by bootstrap) of the estimates here are much higher than the estimates in the left-hand columns; this is not surprising, as the choice probability estimates in Figure 3 show that explorative choices occur with very low probability, leading to imprecision in the estimates of belief-updating rules following such choices.

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 describe how subjects update beliefs following choices of “blue” ( $Y_t = 2$ ).

Table 4: Estimates of learning (belief-updating) rules  
Each cell contains parameter estimates, with bootstrapped standard errors in parentheses. Each column sums to one.

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

$X_t^*$	1( <b>green</b> )	2 (not sure)	3( <b>blue</b> )
$X_{t+1}^* = 1$ ( <b>green</b> )	0.5724 (0.0694)	0.3075 (0.0881)	0.1779 (0.2257)
2 (not sure)	0.0000 <sup>a</sup> (0.0662)	0.3138 (0.1042)	0.4002 (0.2284)
3 ( <b>blue</b> )	0.4276 (0.0624)	0.3787 (0.0945)	0.4219 (0.2195)

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

$X_t^*$	1( <b>green</b> )	2 (not sure)	3( <b>blue</b> )
$X_{t+1}^* = 1$ ( <b>green</b> )	0.8889 (0.0894)	0.6621 (0.1309)	0.8242 (0.2734)
2 (not sure)	0.0000 (0.0911)	0.2702 (0.1297)	0.1758 (0.1981)
3 ( <b>blue</b> )	0.1111 (0.0340)	0.0678 (0.0485)	0.0000 (0.1876)

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

$X_t^*$	3( <b>blue</b> )	2 (not sure)	1( <b>green</b> )
$X_{t+1}^* = 3$ ( <b>blue</b> )	0.5376 (0.0890)	0.2297 (0.0731)	0.2123 (0.1436)
2 (not sure)	0.0458 (0.0732)	0.2096 (0.0958)	0.1086 (0.1524)
1 ( <b>green</b> )	0.4166 (0.0874)	0.5607 (0.0968)	0.6792 (0.1881)

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

$X_t^*$	3( <b>blue</b> )	2 (not sure)	1( <b>green</b> )
$X_{t+1}^* = 3$ ( <b>blue</b> )	0.8845 (0.1000)	0.6163 (0.1136)	0.6319 (0.1647)
2 (not sure)	0.0000 (0.0968)	0.3558 (0.1160)	0.3566 (0.1637)
1 ( <b>green</b> )	0.1155 (0.0499)	0.0279 (0.0373)	0.0116 (0.0679)

<sup>a</sup>This estimate, as well as the other estimates in this table which are equal to zero, resulted from applying the constraint that probabilities must lie between 0 and 1. See the discussion in Section 3 for more details.

## 4 How optimal are estimated learning rules?

In the remainder of the paper, we compare our estimated learning rules to alternative learning rules which have been considered in the literature. We consider four alternative parametric learning rules: (i) the *optimal dynamic Bayesian* model, which is the model discussed in Section 1.1 above; (ii) a *pseudo-Bayesian* model, which is a version of the optimal Bayesian model in which the decision rules are smoothed relative to the step-function decision rules in the optimal model (cf. Figure 2); (iii) *reinforcement learning* (cf. Sutton & Barto (1998)); and (iv) *win-stay*, a simple choice heuristic whereby subjects replay successful strategies. All of these models, except (i), contain unknown model parameters, which we estimated using the choice data from the experiments. Complete details on these models, and the estimated model parameters, are given in Appendix B.

The relative optimality of each learning model was assessed via simulation. For each model, we simulated 100,000 sequences (each containing eight blocks of choices, as in the experiments) of rewards and choices, and computed the distributions of payoffs obtained by agents. The empirical quantiles of these distributions are presented in Table 5.

Table 5: Simulated payoffs from learning models

	Optimal Bayesian <sup>a</sup>	Nonparametric	Pseudo-Bayesian	Reinforcement Learning	Win-stay
5-%tile	\$5	\$1	\$2	\$1	\$1
25-%tile	\$12	\$8	\$9	\$8	\$8
50-%tile	\$17	\$13	\$14	\$13	\$13
75-%tile	\$22	\$18	\$19	\$18	\$18
95-%tile	\$29	\$25	\$26	\$25	\$25

<sup>a</sup>As described in Section 1.1

Reinforcement learning, Pseudo-Bayesian, and win-stay models are described in Appendix B. For each model, the quantiles of the simulated payoff distribution (across 100,000 simulated choice/reward sequences) is reported.

As we expect, the optimal Bayesian model generates the most revenue for subjects; the simulated payoff distribution for this model stochastically dominates the other models, and the median payoff is \$17. The other models perform almost identically, with a median payoff around \$3-\$4 less than the Bayesian model (or about two cents per choice). This difference

accounts for about 25% of typical experimental earnings (not counting the fixed show-up fee).

In the next section, we look for explanations for the differences (and similarities) in performance among the alternative learning models by comparing the belief-updating and choice rules across the different models.

## 4.1 Comparing choice and belief-updating rules across different learning models

For the optimal Bayesian and reinforcement learning models, we can recover the “beliefs” corresponding to the observed choices and rewards, and compare them to the beliefs from the nonparametric learning model.<sup>22</sup> Appendix B contains additional details on how the beliefs were derived for the learning models.

In Table 6, we present some summary statistics for the implied beliefs from our nonparametric learning model (denoted  $X_t^*$ ), vs. the Bayesian beliefs  $B^*$  and the valuations  $V^*$  in the Reinforcement 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 Reinforcement Learning model, and the Bayesian model.<sup>23</sup> 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. Subjects’ beliefs tended to favor green and blue roughly equally, with “not sure” lagging far behind. The close split

---

<sup>22</sup>There are no beliefs in the win-stay model, which is a simple choice heuristic. The pseudo-Bayesian model has the same beliefs as the optimal Bayesian model (with the difference that the choice rule is smoothed).

<sup>23</sup>As we clarify in Appendix B, the nonparametric beliefs  $X_t^*$  were estimated from a maximum likelihood procedure which ignores the implied correlation between choices and beliefs; this is because, given the estimates of the choice probabilities in Table 3, which showed that  $P(Y_t = 1|X_t^* = 1) \approx P(Y_t = 2|X_t^* = 3) \approx 1$ , estimating beliefs  $X_t^*$  based on observed choices  $Y_t$  would lead to estimates of beliefs which practically coincide with choices (ie.  $X_t^* = Y_t$ ), an artificially good “fit” which we felt does not accurately represent the belief process of the subjects.

Table 6: Summary statistics for beliefs in three learning models

$X^*$ : Beliefs from nonparametric model  
 $B^*$ : Beliefs from Bayesian model  
 $V^*$ : “Beliefs” (valuations) from reinforcement learning model **Panel 1:**

$X^*$	1( <span style="color: green;">green</span> )	2(not sure)	3( <span style="color: blue;">blue</span> )
	1878 (45%)	366 (10%)	1956 (45%)

**Panel 2:**

	mean	median	std.	33%-tile	33%-tile
$B^*$ (Bayesian Belief)	0.4960	0.5000	0.1433	0.4201	0.5644
$V^*(= V_b - V_g)$	-0.0104	0	0.4037	-0.2095	0.1694

See Appendix B for details on computation of beliefs in these three learning models.

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 Reinforcement Learning and Bayesian models. The Reinforcement Learning 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”.

Next, we compare the learning rules from the nonparametric, (optimal) Bayesian, and reinforcement learning models. In order to do this, we discretized the beliefs in each model into three values, in proportions identical to the frequency of the different values of  $X_t^*$  as reported in Table 6, and present the implied learning rules for each model.<sup>24</sup> These are shown in Table 7.

Comparing the three sets of learning rules, we see that the most striking difference between

---

<sup>24</sup>Specifically, we discretized the Bayesian (resp. Reinforcement Learning) beliefs so that 45% of the beliefs fell in the  $B_t^* = 1$  (resp.  $V_t^* = 1$ ) and  $B_{t+1}^* = 3$  (resp.  $V_t^* = 3$ ) categories, while 10% fell in the intermediate  $B_t^* = 2$  ( $X_t^* = 2$ ) category, the same as for the nonparametric beliefs  $X_t^*$  (cf. Panel 1 of Table 6). The results are even more striking when we discretized the Bayesian and Reinforcement Learning beliefs so that 33% fell into each of the three categories.

Table 7: Learning (belief-updating) rules for alternative learning models

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

	Optimal Bayesian Learning			Reinforcement Learning		
Beliefs $B_{t+1}^*, V_{t+1}^*$ :	1(green)	2 (not sure)	3(blue)	1(green)	2 (not sure)	3(blue)
1 (green)	0.2878	0	0	0.6538	0	0
2 (not sure)	0.1730	0	0	0.1381	0.0115	0
3 (blue)	0.5392	1.0000	1.0000	0.2080	0.9885	1.0000

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

	Optimal Bayesian Learning			Reinforcement Learning		
Beliefs $B_{t+1}^*, V_{t+1}^*$ :	1(green)	2 (not sure)	3(blue)	1(green)	2 (not sure)	3(blue)
1 (green)	1.0000	1.0000	0.6734	1.0000	0.8818	0.6652
2 (not sure)	0	0	0.1250	0	0.1182	0.1674
3 (blue)	0	0	0.2016	0	0	0.1674

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

	Optimal Bayesian Learning			Reinforcement Learning		
Beliefs $B_{t+1}^*, V_{t+1}^*$ :	3(blue)	2 (not sure)	1(green)	3(blue)	2 (not sure)	1(green)
3 (blue)	0.3060	0	0	0.6576	0	0
2 (not sure)	0.1601	0	0	0.1261	0.0109	0
1 (green)	0.5338	1.0000	1.0000	0.2164	0.9891	1.0000

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

	Optimal Bayesian Learning			Reinforcement Learning		
Beliefs $B_{t+1}^*, V_{t+1}^*$ :	3(blue)	2 (not sure)	1(green)	3(blue)	2 (not sure)	1(green)
3 (blue)	1.0000	1.0000	0.6760	1.0000	0.8898	0.6983
2 (not sure)	0	0.0000	0.1440	0	0.1102	0.1379
1 (green)	0	0	0.1800	0	0	0.1638

them is in how beliefs update following unsuccessful choices (ie. choices which yielded a negative reward). Comparing the Bayesian and the nonparametric learning rules (in Table 5), we see that Bayesian beliefs exhibit less “stickiness”, or serial correlation, following unsuccessful choices. For example, consider the case of  $(Y_t = 1, R_t = 1)$ , so that an unsuccessful choice of green occurred in the previous period. The nonparametric learning rules (Table 5) show that the weight of beliefs remain on “green” ( $X_{t+1}^* = 1$ ) with 57% probability, whereas the Bayesian beliefs place only 28% weight on green. A similar pattern exists after an unsuccessful choice of blue, as shown in the left-hand column of the third panel.

On the other hand, the learning rules for the Reinforcement Learning model (also reported in Table 7) are more similar to the nonparametric learning rule, especially following unsuccessful choices. Again, looking at the top panel, we see that following an unsuccessful choice of “green” ( $Y_t = 1$ ), subjects valuations are still favorable to green with probability 65%; this is comparable in magnitude to the 57% from the nonparametric learning rule. Similarly, after an unsuccessful choice of blue (third panel), valuations in the Reinforcement Learning model still favor blue with probability 66%, again comparable to the 54% for the nonparametric model. It appears that the updating rules from the Reinforcement Learning and nonparametric model share a common defect: a reluctance to “update down” following unsuccessful choices; this common defect relative to the optimal Bayesian model may explain the lower revenue generated by these models.

In Table 8 we compare the choice rules across the different models. As in the previous table, we discretized the beliefs from each model into three values. Comparing the top two panels, we see that, even though the belief-updating rule is the same for the Optimal Bayesian and Pseudo-Bayesian models, the choice rules are strikingly different. Evaluated at the estimated model parameter (discussed in Appendix A), choice probabilities in the Pseudo-Bayesian model are practically invariant to the beliefs, and equal to around 50% for all values of beliefs.

In contrast, choice rules in the Optimal Bayesian model are deterministic functions of beliefs. Overall, the estimated choice rules in Table 3 are much closer to the Optimal Bayesian model, than the Pseudo-Bayesian model. This suggests that the lower payoffs from the estimated

Table 8: Choice rules for alternative learning models

<b>Optimal Bayesian Learning</b>			
Beliefs $B_t^*$ :	1( <b>green</b> )	2(not sure)	3( <b>blue</b> )
$Y_t = 1$ ( <b>green</b> )	1.0000	0.5000	0.0000
2 ( <b>blue</b> )	0.0000	0.5000	1.0000

<b>Pseudo-Bayesian Learning</b>			
Beliefs $B_t^*$ :	1( <b>green</b> )	2(not sure)	3( <b>blue</b> )
$Y_t = 1$ ( <b>green</b> )	0.5141	0.4996	0.4850
2 ( <b>blue</b> )	0.4859	0.5005	0.5150

<b>Reinforcement Learning</b>			
Beliefs $V_t^*$ :	1( <b>green</b> )	2(not sure)	3( <b>blue</b> )
$Y_t = 1$ ( <b>green</b> )	0.7629	0.4939	0.2250
2 ( <b>blue</b> )	0.2371	0.5061	0.7750

model relative to the Optimal Bayesian model arise primarily not from the choice rules (which are very similar in the two models), but rather from the belief-updating rules (which are quite different, as discussed previously).

The bottom panel of Table 8 contains the choice rules for the Reinforcement Learning model. As shown there, the choice rules are much smoother than in the Optimal Bayesian model and the estimated model, but not as smooth as the Pseudo-Bayesian model. This suggests that the similarities of the payoffs from the estimated model relative to Reinforcement Learning (as shown in Table 5) arise mainly from the similarities in belief-updating rules, and less from the choice rules, which are quite different in the two models.

Finally, the similarity in payoffs between the nonparametric and win-stay models is not surprising because, as we showed in Section 1.3 above, the reduced-form choice behavior from the experimental data is in line with a “win-stay/lose-randomize” rule of thumb. Such behavior is confirmed in the formal parameter estimates for the win-stay model (presented in Appendix B.5) which show that, after receiving a positive reward, subjects tend to repeat the previous choice with probability 87% while, after a negative reward, subjects essentially randomize. This asymmetry in choices following good/bad rewards echoes the nonparametric learning rules from Table 5, which showed that subjects “update down” much less following bad rewards than they “update up” following good rewards.

## 4.2 Are eye movements noisy measure of beliefs?

The empirical exercise we undertake in this paper hinges crucially on the assumption that eye movements are constitute (noisy) measurements of the unobserved beliefs, and are not merely noisy measurements of choices. Having used the choice data to estimate beliefs for the nonparametric learning model, as well as the benchmark Bayesian model, we conclude the paper by using these beliefs to perform an assessment of this critical assumption.<sup>25</sup>

Independently of our empirical model and its underlying assumptions, eye movements are really related to some intuitive notion of how beliefs evolve? Using Bayesian beliefs as such an intuitive (and objective) measure of beliefs, we compare each Bayesian belief  $B_{it}^*$  to the corresponding (undiscretized) eye movement measure  $\tilde{Z}_{it}$  (as defined in Eq. (3)) recorded for that subject and trial. The graphs are presented in Figure 3. In the top graph, we see that  $Z$  is clearly increasing with  $B^*$ , suggesting that eye movements track well a standard notion of beliefs.

Of course, this positive relationship could be spurious; if eye movements were not a noisy measure of beliefs, but rather of choice, then the graph here may be picking up simply the common dependence of both  $\tilde{Z}$  and  $B^*$  on choices. To address this, we consider, in the remaining two graphs in Figure 3, a plot of  $(\tilde{Z}_t, B_t^*)$  values *conditional on the choice*  $Y_t$ ; by conditioning on choice, we eliminate any variation in eye movements due to differences in choice.

For the most part, we see that the positive relationship between  $\tilde{Z}_t$  and  $B_t^*$  remains, even after conditioning on choice.<sup>26</sup> For example, the overall positive trend in the second graph suggests, reasurably, that the eye tracking measure  $\tilde{Z}_t$  more strongly favored blue (ie.  $\tilde{Z}_t$  takes large values) when there was strong evidence that blue was the good arm (ie.  $B_t^*$  is large), than when there is only weaker evidence (ie.  $B_t^*$  is small), even after controlling for the relationship between eye movements and choices. This is solid evidence that eye

---

<sup>25</sup>See also Appendix E for another approach to assessing this assumption.

<sup>26</sup>Moreover, the data will be less concentrated in the region of small values of  $B^*$  in the second graph (when blue would tend not to be chosen), and large values of  $B^*$  in the bottom graph (where green would tend not to be chosen). This may explain the kinks in the graphs in those regions.

movements are related to some intuitive notion of how beliefs behave, and are not simply noisy measures of choices.<sup>27</sup>

## 5 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 measurements 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 that subjects tend to update asymmetrically after unsuccessful choices as compared to successful choices. The profits from following the estimated learning and decision rules are smaller than what would be obtained from an optimal Bayesian learning model (about \$4 less for each subject, at the median), and comparable to the profits obtained from three other parametric models: a reinforcement learning model, a “pseudo”-Bayesian model, and a win-stay choice heuristic. Relative to the optimal Bayesian model, the belief-updating rules from the nonparametric and Reinforcement Learning model share a common feature that subjects appear reluctant to “update down” following unsuccessful choices; this may explain the sub-optimality of these models (in terms of profits).

Our nonparametric estimator for subjects’ choice probabilities and learning rules is easy to implement, involving only elementary matrix operations. Furthermore, from a methodological point of view, the *modus operandi* used in this paper – the nonparametric estimation of learning models using experimental data – appears to be a portable idea which can be potentially applied more broadly to other experiments involving dynamic decision problems.

## References

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

---

<sup>27</sup>See also Appendix E for an alternative assessment of this.

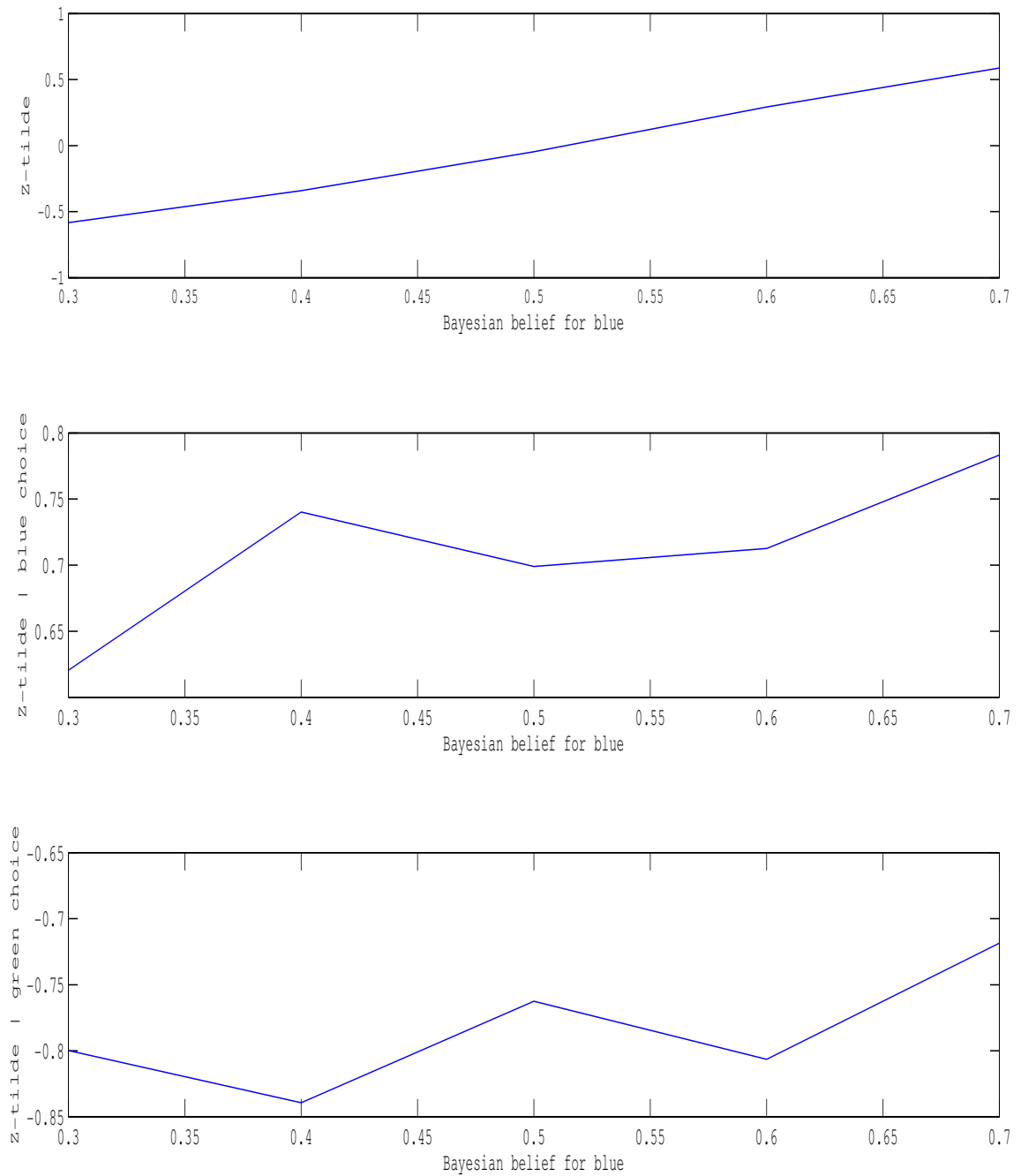


Figure 3: How eye movements track Bayesian beliefs

X-axis: Bayesian beliefs  $B^*$  (detail in Appendix B.2); Y-axis: Undiscretized eye movement measure  $\tilde{Z}$  (defined in Eq. (3)).

- Arcidiacono, P., and R. Miller.** 2006. "CCP Estimation of Dynamic Discrete Choice Models with Unobserved Heterogeneity." Manuscript, Duke University.
- 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.
- Brocas, I., J. Carrillo, S. Wang, and C. Camerer.** 2009. "Measuring attention and strategic behavior in games with private information." mimeo., USC.
- Chan, T., and B. 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.
- 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. J. Gani et. al. North-Holland.
- Grether, D.** 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.
- Hikosaka, O., K. Nakamura, and H. Nakahara.** 2006. “Basal ganglia orient eyes to reward.” *Journal of neurophysiology*, 95(2): 567.
- 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.” Johns 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.
- Kawagoe, R., Y. Takikawa, and O. Hikosaka.** 1998. “Expectation of reward modulates cognitive signals in the basal ganglia.” *Nat Neurosci*, 1: 411–416.
- Krajbich, I., C. Armel, and A. Rangel.** 2010. “Visual fixations and the computation and comparison of value in simple choice.” *Nature Neuroscience*, 13: 1292–1298.
- Kuhnen, C., and B. Knutson.** 2008. “The Influence of Affect on Beliefs, Preferences and Financial Decisions.” University Library of Munich, Germany MPRA Paper 10410.
- Lauwereyns, J., K. Watanabe, B. Coe, and O. Hikosaka.** 2002. “A neural correlate of response bias in monkey caudate nucleus.” *Nature*, 418: 413–417.
- 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.
- Miller, R.** 1984. “Job Matching and Occupational Choice.” *Journal of Political Economy*, 92: 1086–1120.
- Nyarko, Y., and A. Schotter.** 2002. “An Experimental Study of Belief Learning Using Elicited Beliefs.” *Econometrica*, 70: 971–1005.
- Odean, T., M. Strahilevitz, and B. Barber.** 2004. “Once Burned, Twice Shy: How Naive Learning and Counterfactuals Affect the Repurchase of Stocks Previously Sold.” 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-LeNestour, E., and P. Bossaerts.** 2011. “Risk, Unexpected Uncertainty, and Estimation Uncertainty: Bayesian Learning in Unstable Settings.” *PLoS Computational Biology*, 7(1): 1704–1711.
- 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.

- Rescorla, R., and A. Wagner.** 1972. "Variations in the Effectiveness of Reinforcement and Non-reinforcement." *New York: Classical Conditioning II: Current Research and Theory, Appleton-Century-Crofts.*
- Reutskaja, E., R. Nagel, C.F. Camerer, and A. Rangel.** 2011. "Search Dynamics in Consumer Choice under Time Pressure: An Eye-Tracking Study." *American Economic Review*, 101(2): 900–926.
- 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.
- Shimojo, S., C. Simion, E. Shimojo, and C. Scheier.** 2003. "Gaze bias both reflects and influences preference." *Nature neuroscience*, 6(12): 1317–1322.
- Sutton, R., and A. Barto.** 1998. *Reinforcement Learning*. MIT Press.
- Wang, J.T., M. Spezio, and C.F. Camerer.** 2010. "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games." *American Economic Review*, 100(3): 984–1007.
- Yoshida, W., and S. Ishii.** 2006. "Resolution of uncertainty in prefrontal cortex." *Neuron*, 50(5): 781–789.
- Zhang, J., M. Wedel, and R. Pieters.** 2009. "Sales Effects of Attention to Feature Advertisements: a Bayesian Mediation Analysis." *Journal of Marketing Research*, 46: 669–681.

## Supplemental appendices: not for publication

### A Details of optimal Bayesian learning model

Here we provide more details about the simulation of the optimal learning and decision rules from Section 1.1. First we introduce some notation and describe the information structure and how Bayesian updating would proceed in the reversal learning context. Let  $(Y_t, S_t, R_t)$  denote the actions, state, and rewards. Furthermore, let  $Q$  denote the  $2 \times 2$  Markov transition matrix for the state  $S_t$ , corresponding to Eq. (2).

Let  $B_t^*$  denote the *prior belief* that  $S_t = 1$ , at the beginning of period  $t$ , while  $\tilde{B}_t^*$  denotes the *posterior belief* that  $S_t = 1$ , at the end of period  $t$ , after taking action  $Y_t$  and observing reward  $R_t$ . The relationship between  $B_t^*$  and  $\tilde{B}_t^*$  is given by Baye's rule:

$$\tilde{p}_t = P(S_t = 1 | p_t, R_t, Y_t) = \frac{p_t \cdot f(R_t | S_t = 1, Y_t)}{p_t \cdot f(R_t | S_t = 1, Y_t) + (1 - p_t) \cdot f(R_t | S_t = 2, Y_t)}$$

Combining this with  $Q$ , we obtain the period-by-period transition for the prior beliefs  $B_t^*$ :

$$\begin{bmatrix} B_{t+1}^* \\ 1 - B_{t+1}^* \end{bmatrix} = Q \cdot \begin{bmatrix} \tilde{B}_t^* \\ 1 - \tilde{B}_t^* \end{bmatrix} = Q \cdot \begin{bmatrix} P(S_t = 1 | B_t^*, R_t, Y_t) \\ 1 - P(S_t = 1 | B_t^*, R_t, Y_t) \end{bmatrix} \quad (10)$$

Next we describe a dynamic Bayesian learning model for the reversal-learning environment. As in the experiments, we consider a finite (25 period) horizon, with  $t = 1, \dots, T = 25$ . Each subject's objective is to choose sequence of actions to maximize expected rewards:

$$\max_{i_1, i_2, \dots, i_T} \mathbb{E} \left[ \sum_{t=1}^T R_t \right]$$

The state variable in this model is  $B_t^*$ , the beliefs at the beginning of each period. Correspondingly, the Bellman equation is:

$$\begin{aligned} V_t(B_t^*) &= \max_{Y_t \in \{1,2\}} \left\{ \mathbb{E} [R_t + V_{t+1}(B_{t+1}^*) | Y_t, B_t^*] \right\} \\ &= \max_{Y_t \in \{1,2\}} \left\{ \mathbb{E} [R_t | Y_t, B_t^*] + \mathbb{E}_{R_t | Y_t, B_t^*} \mathbb{E}_{B_{t+1}^* | B_t^*, Y_t, R_t} V_{t+1}(B_{t+1}^*) \right\} \end{aligned} \quad (11)$$

Above, the expectation  $E_{B_{t+1}^* | B_t^*, Y_t, R_t}$  is taken with respect to Eq. (10), the law of motion for the prior beliefs, while the expectation  $E_{R_t | Y_t, B_t^*}$  is derived from the assumed distribution of

$(R_t|Y_t, \omega_t)$  via

$$P(R_t|Y_t, B_t^*) = B_t^* \cdot P(R_t|Y_t, \omega_t = 1) + (1 - B_t^*) \cdot P(R_t|Y_t, \omega_t = 2).$$

While we have not been able to derive closed-form solutions to this dynamic optimization problem, we can compute the optimal decision rules by backward induction. Specifically, in the last period  $T = 25$ , the Bellman equation is:

$$V_T(B_T^*) = \max_{Y_t \in \{1,2\}} E[R_t|Y_t, B_T^*]. \quad (12)$$

We can discretize the values of  $B_T^*$  into the finite discrete set  $\mathcal{B}$ . Then for each  $B \in \mathcal{B}$ , we can solve Eq. (12) to obtain the period- $T$  value and choice functions  $\hat{V}_T(B)$  and  $\hat{y}_T^*(B) = \operatorname{argmax}_i \mathbb{E}[R_t|i, B]$  for each value of  $B \in \mathcal{B}$ . Subsequently, proceeding backwards, we can obtain the value and choice functions for periods  $t = T - 1, T - 2, \dots, 1$ . These choice functions are plotted in Figure 2.

## B Details on model fitting and belief estimation in different learning models

In section 4, 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.

### B.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  $P(\{X_t^*\} | \{Z_t, R_t\})$ , the conditional (“posterior”) probability of the beliefs, given the observed sequences of 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:

$$\log L(\{X_t^*, Z_t\} | \{Y_t, R_t\}) = \sum_{t=1}^{24} \log [P(Z_t | X_t^*) P(X_{t+1}^* | X_t^*, R_t, Y_t)] + \log(P(Z_{25} | X_{25}^*)). \quad (13)$$

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 (13) above. By doing this, we essentially ignore the implied correlation between beliefs and choices in estimating beliefs. 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 3, maximizing with respect to these choice probabilities would lead to estimates of beliefs  $\{X^*\}$  which closely coincide with observed choices; we wished to avoid such an artificially good "fit" between the beliefs and observed choices.

For robustness, however, we also estimated the beliefs  $\{X^*\}$  including the choice probabilities  $P(Y_t | X_t^*)$  in the likelihood function. Not surprisingly, the correlation between choices and beliefs  $\text{Corr}(Y_t, X_t^*) = 0.99$ , and in practically all periods, the estimated beliefs and observed choices coincided (ie.  $X_t^* = Y_t$ ). However, we felt that this did not accurately reflect subjects' beliefs.

## B.2 Bayesian Learning Model

The learning and decision rules for the Bayesian model were described and computed in Section 1.1, with additional details provided in Appendix A. The sequence of Bayesian beliefs  $B_t^*$  is obtained from Eq. (10) and evaluated at the observed sequence of choices and rewards  $(Y_t, R_t)$ .

### B.3 Reinforcement Learning Model

We employ a variant of the TD (Temporal-Difference)-Learning models (Sutton & Barto (1998) , section 6) in which action values are up-dated via the Rescorla-Wagner rule (R. Rescorla & A. Wagner (1972)). The value updating rule for a one-step TD-Learning model is given by:

$$V_{Y_t}^{t+1} \leftarrow V_{Y_t}^t + \alpha \delta_t. \quad (14)$$

where  $Y_t$  denotes the choice taken in trial  $t$ ,  $\alpha$  denotes the learning rate, and  $\delta_t$  denotes the “prediction error”  $\delta_t$  for trial  $t$ , defined as:

$$\delta_t = R_t - V_{Y_t}^t, \quad (15)$$

the difference between  $R_t$  (the observed reward in trial  $t$ ) and  $V_{Y_t}^t$  (the current valuation). In trial  $t$ , only the value for the chosen alternative  $Y_t$  is updated; there is no updating of the valuation for the choice that was not taken.

$P_c^t$ , the current probability of choosing action  $c$ , is assumed to take the conventional “soft-max” (ie. logit) form with the temperature parameter  $\tau$ :

$$P_c^t = e^{V_c^t/\tau} / \left[ \sum_{c'} e^{V_{c'}^t/\tau} \right] \quad (16)$$

We estimated the parameters  $\tau$  and  $\alpha$  using maximum likelihood. For greater model flexibility, we allowed the parameter  $\alpha$  to differ following positive vs. negative rewards. The estimates (and standard errors) are:

$$\begin{aligned} \tau &= 0.2729 \quad (0.0307) \\ \alpha \text{ for positive reward } (R_t = 2) &= 0.7549 \quad (0.0758) \\ \alpha \text{ for negative reward } (R_t = 1) &= 0.3333 \quad (0.0518). \end{aligned} \quad (17)$$

We plug in these values into Eqs. (14), (15), and (19) to derive a sequence of valuations  $\{V_t^* \equiv V_b^t - V_g^t\}$ . The choice function (Eq. (16)) can be rewritten as a function of the difference  $V_t^*$ ; i.e. the choice probability for the blue slot machine is,

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

and  $P_g^t = 1 - P_b^t$ . Hence,  $V_t^*$  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.4 Pseudo Bayesian Learning Model

A Pseudo Bayesian learner uses Bayes rule to update her belief (as in the Optimal Bayesian model), but her choices are determined (suboptimally) by the "softmax" rule, as in reinforcement learning:

$$P_c^t = e^{B_c^{*t}/\tau} / \left[ \sum_{c'} e^{B_{c'}^{*t}/\tau} \right] \quad (19)$$

The maximum-likelihood estimate of  $\tau$  is 0.2176 with bootstrapped standard error of 0.0138.

## B.5 Win-Stay Model

The final model is a simple behavioral heuristic. If subjects choose a slot machine and receive the positive reward  $R_t = 1$ , they repeat the choice in the next period with probability  $1 - \delta$  (and switch to the other choice with probability  $\delta$ ). If they choose a slot machine but obtain the negative reward  $R_t = -1$ , they switch to the other slot machine in the next trial with probability  $1 - \epsilon$ .

We estimated the parameters  $\delta$  and  $\epsilon$  using maximum likelihood. The estimates we obtained from the data were:

$$\delta = 0.1268 \quad (0.0142); \quad \epsilon = 0.4994 \quad (0.0213). \quad (20)$$

## C Details on discretization of eye movements

In this section, we present additional discussion on the discretization of the eye-movement measure, and some evidence that a three-valued discretization (which we used in our preferred empirical specifications) is sufficient to capture most of the variation in this measure.

We start by assessing the discretization of  $Z_t$  using a statistical approach based on the condition number of the matrix containing the sample conditional probabilities of the discretized values of  $(Z_t|Z_{t-1})$ . The intuition is straightforward: if we discretize into an excessive num-

ber of points, the matrix containing the discretized conditional distribution of  $(Z_t|Z_{t-1})$  will be singular, reflecting the redundancy of information in the overly-discretized values of  $Z_t$ .

Furthermore, and more importantly, our identification assumptions imply that the rank of the matrix  $G_{z_t|z_{t-1}}$  is related to the dimension of the unobserved beliefs  $X^*$ ; thus, a proper discretization of  $Z_t$  is required to obtain reasonable estimates of beliefs. Formally, given a discretization of  $Z$  into  $K$  categories, the largest possible rank is

$$\text{rank}(G_{z_t|z_{t-1}}) = \begin{cases} K & \text{if } K \leq \dim(X^*) \\ \dim(X^*) & \text{if } K > \dim(X^*) \end{cases}.$$

In other words,

$$G_{z_t|z_{t-1}} \text{ is } \begin{cases} \text{nonsingular} & \text{if } K \leq \dim(X^*) \\ \text{singular} & \text{if } K > \dim(X^*) \end{cases}.$$

Therefore, the dimension of  $X^*$  is the largest discretization  $K$  with which  $G_{z_t|z_{t-1}}$  is still nonsingular.

The condition number of a matrix measures how “close to singular” a given matrix is, with a larger condition number indicating a less well-behaved matrix.<sup>28</sup> In this exercise, we discretize  $Z_t$  into number of points ranging from 2 to 6, and computed the condition number for the matrix of sample conditional probabilities  $G(Z_t|Z_{t-1})$  in each case. Table 9 shows the condition numbers for each case, along with block-bootstrap estimates of the percentiles of their sampling distribution.

As the results show, the big jump in condition number occurs between 3 and 4; the sample condition number jumps almost five-fold from 10.7 to 48.1, while the estimated sampling distribution also blows up, with the 95th percentile jumping from 15.2 to 345.9. This offers some statistical confirmation for the three-point discretization of the eye movement measure  $Z_t$  used in our empirical analysis.

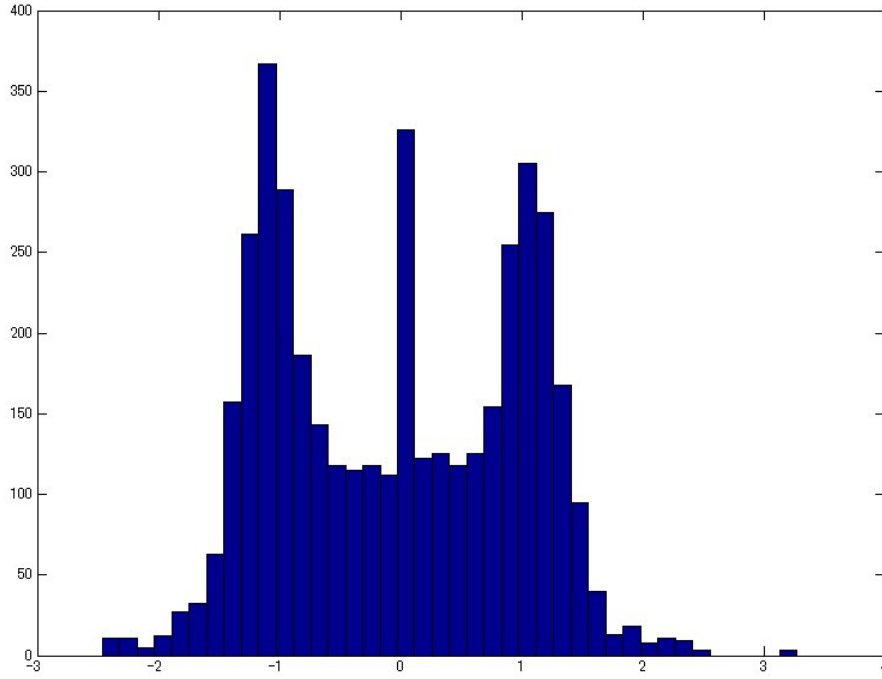
Besides this formal statistical evidence, we also present the raw histogram of the undiscretized

---

<sup>28</sup>Formally, the condition number of a matrix measures the sensitivity of the solution of a system of linear equations to errors in the data, which depends on the invertibility of the matrix containing the coefficients of the linear equations. Values of condition number near 1 indicate a well-conditioned matrix, while a larger condition numbers suggests that a matrix is close to singular.

Table 9: Condition Number of Matrix  $G_{z_t|z_{t-1}}$ 

	Dimension				
	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
original sample	3.3586	10.6571	48.1296	292.3680	198.9212
mean	3.3627	10.7541	176.1842	981.4516	1302.2
minimum	3.2270	6.3758	14.3674	29.7191	30.9333
5th percentile	3.2270	8.2144	22.2867	66.7	62.1
25th percentile	3.2594	9.5753	33.9472	107.9	109.6
median	3.2980	10.7541	49.7292	175.1	185.4
75th percentile	3.3898	12.1294	88.8227	401.6	407.5
95th percentile	3.6331	15.1768	345.9043	1985.0	2530.2
maximum	3.6331	22.6323	39013	273180	222780

Figure 4: Histogram of undiscretized eye movement measure  $\tilde{Z}_t$ 

eye movement measure  $\tilde{Z}_t$  in Figure 4. It is apparently trimodal, with peaks at -1, 0 and 1, suggesting that a three-value discretization of  $Z_p$  indeed captures most of its variation. In the empirical work, we use the following three-value discretization as follows:

$$Z_t = \begin{cases} 1 & \text{if } \tilde{Z}_t < -\sigma_z \\ 2 & \text{if } -\sigma_z \leq \tilde{Z}_t \leq \sigma_z \\ 3 & \text{if } \sigma_z < \tilde{Z}_t \end{cases} \quad (21)$$

where  $\sigma_z$  denotes a discretizing constant. As the baseline, we set  $\sigma_z = 0.20$ . However, we do not find any difference in the estimation results either qualitatively nor significantly if we vary  $\sigma_z$  from 0.05 to around 0.40, suggesting that the model is robust for different classifications. Table 10 shows the sample frequencies of the discretized measure  $Z_t$  for three different values of  $\sigma_z$ .

Moreover, Table 10 also shows the correlations between  $Y$  and  $\tilde{Z}$ , broken up into the three ranges of  $\tilde{Z}$  corresponding to the three discretized values  $Z \in \{1, 2, 3\}$ . Although the correlation between  $(Y, \tilde{Z})$  in the whole sample is 0.7647, the correlations within each of the three ranges of  $\tilde{Z}$  drop significantly, ranging from even negative values to values around 0.30. 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.

## D Conditional serial correlation in eye movements

In this section we assess more formally one critical part of Assumption 3, which is that the eye movement measures are serially independent, conditional on beliefs. That is,  $P(Z_t|Z_{t-1}, X_t^*) = P(Z_t|X_t^*)$ . Since this exclusion restriction plays a crucial role in pinning down the values of the beliefs, we assess it by estimating an alternative model in which we do not impose this assumption. In this alternative model, the “measurement probabilities” are given by the conditional distribution of  $f(Z_t|X_t^*, Z_{t-1})$ . In the remainder of this section, we describe how this expanded model is estimated.

Consider the joint density  $f(Z_t, Y_t, Z_{t-1}, Z_{t-2})$ , which is solely a function of variables observed in the data. Following the approach taken in Section 2.1 of the main text, we can factor this

Table 10: Correlations between  $(Y, \tilde{Z})$  in different subsamples

	Size	Corr( $Y, \tilde{Z}$ )
Full sample	4200	0.7647
<b><math>\sigma_z = 0.20</math> (baseline):</b>		
$Z = 1$ (green)	1887	0.2845
2 (not sure)	540	0.2156
3 (blue)	1773	0.1706
<b><math>\sigma_z = 0.05</math>:</b>		
$Z = 1$ (green)	2015	0.3223
2 (not sure)	255	-0.0599
3 (blue)	1930	0.2346
<b><math>\sigma_z = 0.40</math>:</b>		
$Z = 1$ (green)	1725	0.1462
2 (not sure)	869	0.2777
3 (blue)	1606	0.0991

Note:  $\tilde{Z}_t$  refers to the undiscretized eye-movement measure, as defined in Eq. (3), and  $Z$  refers to the discretized version, as defined in Eq. (21).

density as follows:

$$\begin{aligned}
& f(Z_t, Y_t | Z_{t-1}, Z_{t-2}) \\
&= \sum_{X_t^*} \sum_{X_{t-1}^*} f(Z_t, Y_t, X_t^*, X_{t-1}^* | Z_{t-1}, Z_{t-2}) \\
&= \sum_{X_t^*} \sum_{X_{t-1}^*} f(Z_t | Y_t, X_t^*, X_{t-1}^*, Z_{t-1}, Z_{t-2}) f(Y_t | X_t^*, X_{t-1}^*, Z_{t-1}, Z_{t-2}) f(X_t^*, X_{t-1}^* | Z_{t-1}, Z_{t-2}) \\
&= \sum_{X_t^*} f(Z_t | X_t^*, Z_{t-1}) f(Y_t | X_t^*) \sum_{X_{t-1}^*} f(X_t^*, X_{t-1}^* | Z_{t-1}, Z_{t-2})
\end{aligned}$$

For a fixed  $z_{t-1}$ , we have

$$f(Z_t, Y_t | z_{t-1}, Z_{t-2}) = \sum_{X_t^*} f(Z_t | X_t^*, z_{t-1}) f(Y_t | X_t^*) f(X_t^* | z_{t-1}, Z_{t-2}).$$

Technically, for any fixed  $Y_t = y_t$  and  $Z_{t-1} = z_{t-1}$ , then, we can write the above in matrix notation as:

$$\mathbf{A}_{y_t, Z_t | z_{t-1}, Z_{t-2}} = \mathbf{B}_{Z_t | X_t^*, z_{t-1}} \mathbf{D}_{y_t | X_t^*} \mathbf{C}_{X_t^* | z_{t-1}, Z_{t-2}}$$

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

$$\begin{aligned}
\mathbf{A}_{y_t, Z_t | z_{t-1}, Z_{t-2}} &= [f_{Y_t, Z_t | Z_{t-1}, Z_{t-2}}(y_t, i | z_{t-1}, j)]_{i,j} \\
\mathbf{B}_{Z_t | X_t^*, z_{t-1}} &= [f_{Z_t | X_t^*, Z_{t-1}}(i | k, z_{t-1})]_{i,k} \\
\mathbf{C}_{X_t^* | z_{t-1}, Z_{t-2}} &= [f_{X_t^* | Z_{t-1}, Z_{t-2}}(k | z_{t-1}, j)]_{k,j}
\end{aligned}$$

The key eigendecomposition equation, analogous to Eq. (7) for the simpler model, becomes

$$\mathbf{A}_{y_t, Z_t | z_{t-1}, Z_{t-2}} \mathbf{G}_{Z_t | z_{t-1}, Z_{t-2}}^{-1} = \mathbf{B}_{Z_t | X_t^*, z_{t-1}} \mathbf{D}_{y_t | X_t^*} \mathbf{B}_{Z_t | X_t^*, z_{t-1}}^{-1}$$

where<sup>29</sup>

$$\mathbf{G}_{Z_t | z_{t-1}, Z_{t-2}} = [f_{Z_t | Z_{t-1}, Z_{t-2}}(i | z_{t-1}, k)]_{i,k}$$

Therefore, we can apply this eigendecomposition to estimate  $f(Z_t | X_t^*, z_{t-1})$  for each value of  $z_{t-1}$ . To assess whether we need to allow for conditional serial correlation in  $Z_t$ , we can compare whether the estimated probabilities  $f(Z_t | X_t^*, z_{t-1})$  differ in  $z_{t-1}$ , i.e.

$$f(Z_t | X_t^*, \tilde{z}_{t-1}) \stackrel{?}{=} f(Z_t | X_t^*, \bar{z}_{t-1}).$$

<sup>29</sup>Note that the invertibility of  $\mathbf{G}_{Z_t | z_{t-1}, Z_{t-2}}$  is testable for each  $z_{t-1}$ .

Table 11: Measurement probabilities:  $P(Z_t|X_t^*)$ 

		$Z_{t-1} = 1, (N = 1748)$		
		$X_t^* = 1$	$X_t^* = 2$	$X_t^* = 3$
$Z_t = 1$		0.8522 (0.1060)	0.2138 (0.1508)	0.0797 (0.0472)
$Z_t = 2$		0.0923 (0.0649)	0.4523 (0.1264)	0.1257 (0.0508)
$Z_t = 3$		0.0555 (0.0546)	0.3340 (0.1285)	0.7945 (0.0679)
		$Z_{t-1} = 2, (N = 487)$ Insufficient sample size.		
		$Z_{t-1} = 3, (N = 1629)$		
		$X_t^* = 1$	$X_t^* = 2$	$X_t^* = 3$
$Z_t = 1$		0.7844 (0.0950)	0.1706 (0.1170)	0.0574 (0.0513)
$Z_t = 2$		0.0732 (0.0553)	0.5398 (0.2023)	0.1744 (0.1160)
$Z_t = 3$		0.1425 (0.0697)	0.2879 (0.2019)	0.7682 (0.1378)

note 1: cut-off for the three-value discretization is 0.2

note 2: Standard errors (in parentheses) computed across 1500 bootstrap resamples

The estimates of the probabilities  $f(Z_t|X_t^*, z_{t-1})$  for  $z_{t-1} = 1, 3$  are presented in Table 11.<sup>30</sup> As the results show, the estimates of these probabilities are quite similar across different values of  $z_{t-1}$ . This suggests that conditional serial correlation in eye movements is not a major concern, and supports Assumption 3 underlying our empirical model.

<sup>30</sup>We were not able to estimate  $f(Z_t|X_t^*, z_{t-1} = 2)$  because we observed too few observations with  $z_{t-1} = 2$ .

## E Belief-updating and choices following “unsure” belief state

In section 4.2 of the main text, we present some evidence, based on comparing the eye movement measures to the beliefs from the Bayesian model, that eye movements were noisy measurements of beliefs, and not just of choices. Here, we consider another assessment of this crucial assumption which underlies our empirical model.

Here, we exploit that fact that, in our model, beliefs (and eye movements) take more values than choices. We consider what happens when beliefs are “unsure”; that is, when beliefs  $X_t^*$  take the intermediate value of 2. Since eye movements play a crucial role in pinning down beliefs, if eye movements are just a noisy measure of choices, then choices should be similar following the “unsure” state ( $X_t^* = 2$ ) than following the “sure” states ( $X_t^* = 1$  or  $X_t^* = 3$ ). However, if eye movements contain extra information beyond that contained in the choices, then we should find that belief updating and choice behavior following the “unsure” state is different than that following the “sure” states. The goal is to show that the “unsure” state matters for both belief updating and decision-making.<sup>31</sup>

First we show that beliefs update different following the unsure state than following a sure state. To do this, we perform a joint test that the probabilities in the left-most column (corresponding to beliefs following belief-congruent choices) of each transition matrix in Table 4 differs from the middle column. We construct the test statistic as follows. Let  $\vec{L}$  (resp.  $\vec{M}$ ) denote the left-hand (resp. middle) column of a matrix in Table 4, omitting the bottom element. The test statistic is the quadratic form  $(\vec{L} - \vec{M})' \Sigma^{-1} (\vec{L} - \vec{M})$ , where  $\Sigma$  is the variance-covariance matrix of  $(\vec{L} - \vec{M})$  which was computed by bootstrap (as was all the estimates in Table 4).

Asymptotically, under the null hypothesis of no differences between the columns, this statistic is distributed according to a  $\chi^2$ -distribution, with two degrees of freedom. The corresponding  $p$ -values are given in Table 12. The  $p$ -values are all small; the first two  $p$ -values imply that

---

<sup>31</sup>We are grateful to a referee for this suggestion.

the “not sure” and “green” belief states are distinct, while the last two indicate that the “not sure” and “blue” states differ. Hence, the unsure state ( $X_t^* = 2$ ) matters, in the sense that beliefs in the following period  $X_{t+1}^*$  are statistically different when  $X_t^*$  is a sure (“blue”, “green”) state versus an unsure state.

Table 12: Tests of belief updating following unsure state

$$H_0: P(X_{t+1}^* | X_t^* = 1, Y_t, R_t) = P(X_{t+1}^* | X_t^* = 2, Y_t, R_t)$$

$(R_t, Y_t):$	(1, 1)	(2, 1)	(1, 2)	(2, 2)
$p$ -value: LH=middle <sup>a</sup>	0.016	0.121	0.032	0.072

<sup>a</sup>Each entry contains the  $p$ -value under the null hypothesis that the leftmost and middle columns in the corresponding matrix in Table 5 have the same values. Under the null hypothesis, the test statistic has an asymptotic  $\chi^2$  distribution with two degrees of freedom.

However, beliefs are unobservable. How does the unsure state affect future *observed* choices?

To do this, we used our estimation results to compute the conditional distributions

$$Y_{t+1} | X_t^*, Y_t, R_t = \sum_{i=1}^3 (Y_{t+1} | X_{t+1}^* = i) \cdot (X_{t+1}^* = i | X_t^*, Y_t, R_t).$$

The conditional distribution  $Y_{t+1} | X_t^*, Y_t, R_t$  describes how choices in period  $t$  are made, conditional of beliefs, choices, and rewards in period  $t$ . This distribution is given in Table 13. As before, we want to test whether the “unsure” state ( $X_t^* = 2$ ) has distinctive effects on observed choices. To do this, we test, as before, whether the leftmost and middle columns of each matrix in Table 13 are the same. The  $p$ -values under the null that these two columns are the same are also reported in Table 13. These  $p$ -values are small, indicating scant evidence favoring the null hypothesis; while we cannot reject the null hypothesis at conventional significance levels in two of the four cases (corresponding to  $(Y_t = 1, R_t = 1)$  and  $(Y_t = 2, R_t = 2)$ ), the small  $p$ -values do favor the hypothesis that the leftmost and middle columns are different. Thus, we also find that the “unsure” state is important in predicting choices, which implies that the eye movements  $Z_t$  contain more information than is contained in choices alone. This lends support to our modelling assumption that eye movements are noisy measures of beliefs.

Table 13: How current beliefs affect future choices  
 The conditional probabilities  $Y_{t+1}|X_t^*, Y_t, R_t$  computed from estimation results.

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

$X_t^*$	1(green)	2 (not sure)	3(blue)
$Y_{t+1} = 1$ (green)	0.5675	0.4445	0.3551
2 (blue)	0.4325	0.5555	0.6449

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

$X_t^*$	1(green)	2 (not sure)	3(blue)
$Y_{t+1} = 1$ (green)	0.8777	0.7731	0.8909
2 (blue)	0.1223	0.2270	0.1091

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

$X_t^*$	3(blue)	2 (not sure)	1(green)
$Y_{t+1} = 2$ (blue)	0.5653	0.3527	0.2806
1 (green)	0.4347	0.6473	0.7195

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

$X_t^*$	3(blue)	2 (not sure)	1(green)
$Y_{t+1} = 2$ (blue)	0.8795	0.8114	0.8270
1 (green)	0.1205	0.1980	0.1731

$$H_0: P(Y_{t+1}|X_t^* = 1, Y_t, R_t) = P(Y_{t+1}|X_t^* = 2, Y_t, R_t)$$

$(Y_t, R_t)$ :	(1, 1)	(1, 2)	(2, 1)	(2, 2)
$p$ -value: LH=middle <sup>a</sup>	0.109	0.091	0.022	0.163

<sup>a</sup>Each entry contains the  $p$ -value under the null hypothesis that the leftmost and middle columns in the corresponding matrix in the above table have the same values. Under the null hypothesis, the test statistic has an asymptotic standard normal distribution.

## F Additional figures

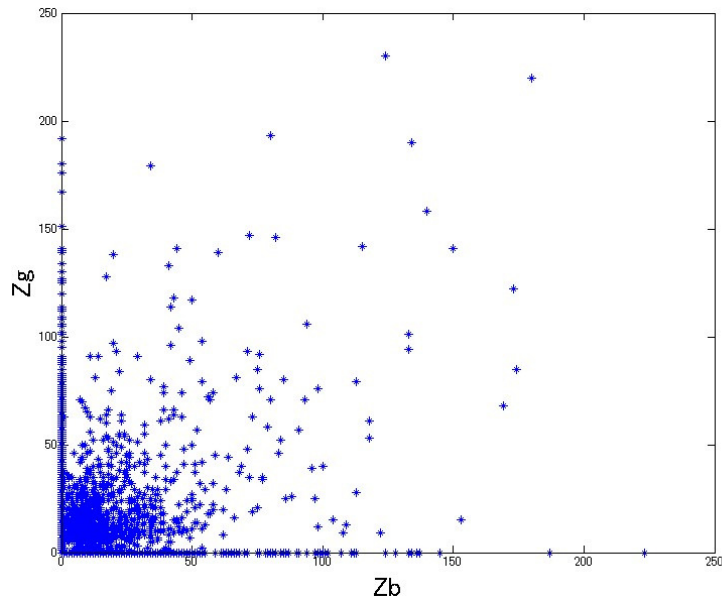


Figure 5: Scatter plot of  $Z_b$  (fixation on blue) and  $Z_g$ (fixation on green)

Both  $Z_b$  and  $Z_g$  are reported in  $2 \times 10^{-2}$  seconds.