DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES

CALIFORNIA INSTITUTE OF TECHNOLOGY

PASADENA, CALIFORNIA 91125

NONPARAMETRIC LEARNING RULES FROM BANDIT EXPERIMENTS: THE EYES HAVE IT!

Yingyao Hu
Johns Hopkins University

Yutaka Kayaba
Caltech

Matthew Shum
Caltech

SOCIAL SCIENCE WORKING PAPER 1326

June 2010

Nonparametric Learning Rules from Bandit Experiments:
The Eyes have it!

Yingyao Hu Yutaka Kayaba Matthew Shum

Abstract

How do people learn? We assess, in a distribution-free manner, subjects' learning and choice rules in dynamic two-armed bandit learning experiments. 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 that subjects tend to update in a non-smooth manner following choices made in accordance with current beliefs. Moreover, the beliefs implied by our nonparametric learning rules are closer to those from a (non-Bayesian) reinforcement learning model, than a Bayesian learning model.

JEL classification numbers: D83, C91, C14

Key words: learning, experiments, eye-tracking, Bayesian vs. non-Bayesian learning, nonparametric estimation

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

Yingyao Hu Yutaka Kayaba Matthew Shum

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 experiments, 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 nonparametrically allows us to compare competing learning models 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 Miller (1984), Erdem and Keane (1996), Ackerberg (2003), Crawford and Shum (2005), Chan and Hamilton (2006), and Marcoul and Weninger (2008). This literature typically assumes that agents process information according to a forward-looking Bayesian learning model. This restrictive assumption is driven in part by data considerations: oftentimes, all that is observed are the sequences of agents' choices, so that a lot of (parametric) structure must be placed on the learning model for identification.

*Hu: yhu@jhu.edu. Kayaba & Shum: {ykayaba,mshum}@caltech.edu. We are indebted to Antonio Rangel for his encouragement and for the funding and use of facilities in his lab. We thank Dan Ackerberg, Peter Bossaerts, Colin Camerer, Andrew Ching, Cary Frydman, Ian Krajbich, Pietro Ortoleva, and participants in presentations at U. Arizona, Caltech, UCLA, U. Washington and Choice Symposium 2010 (Key Largo) for comments and suggestions.

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 brain activity (cf. Yoshida and Ishii (2006), Boorman, Behrens, Woolrich, and Rushworth (2009) in the recent fMRI neuroscience literature), mouse-tracking (cf. Brocas, Carrillo, Wang, and Camerer (2009)), or eye movements (as in Armel and Rangel (2008), and the present paper).

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 "reinforcement learning" (RL) rules (cf. Sutton and Barto (1998)). An incomplete list of papers which consider these questions includes Grether (1992), El-Gamal and Grether (1995), Charness and Levin (2005), Kuhnen and Knutson (2008), and Payzan and Bossaerts (2009). Particularly, RL has attracted considerable attention in the recent neuroeconomics and decision neuroscience literature (cf. Glimcher, Camerer, Poldrack, and Fehr (2008), Rushworth and Behrens (2008)), ever since studies showing that the "prediction errors" of these models are apparently encoded in certain areas of the brain (cf. Schultz, Dayan, and Montague (1997)) for evidence from primates). Recently, RL models have also been used to explain some observed anomalies in savings and investment behavior (eg. Choi, Laibson, Madrian, and Metrick (2009), Odean, Strahilevitz, and Barber (2004)).[1]

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

Moreover, we estimate not only the learning rules nonparametrically, but also the choice probabilities. Choice probabilities are key parameters in machine learning and de-

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

cision neuroscience models (cf. Sutton and Barto (1998), Doya (2002)). Although several studies have examined parameterized models of choice behavior (cf. Daw, O'Doherty, Dayan, Seymour, and Dolan (2006)), to our knowledge, the present paper is the first which examines choice behavior in learning models without imposing *a priori* functional forms on the choice probabilities.

Our approach differs from a common *modus operandi* in the behavioral/experimental literature, which has been to use the observed choice data from the experiment to calibrate parameters for competing learning models. Subsequently, the competing learning models are simulated, and verification is based upon comparing the simulated learning rules with the observed auxiliary belief measurements. For instance, Hampton, Bossaerts, and O'Doherty (2006) test between a Bayesian and reinforcement-learning model on the basis of two-armed bandit experiments supplemented with brain activity information from fMRI brainscans.[2] 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 (Hu (2008), Hu and Shum (2008)). Because subjects' underlying beliefs are unobserved and also serially correlated over time, the learning model is a particular case of a nonlinear "hidden state Markov" model, which can be challenging to estimate.[3] In contrast, by fitting 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 which involves only elementary calculations involving matrices which can be formed from the observed data.[4]

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.

[2]Other papers utilizing a similar methodological framework include Behrens, Woolrich, Walton, and Rushworth (2007), Boorman, Behrens, Woolrich, and Rushworth (2009), Daw, O'Doherty, Dayan, Seymour, and Dolan (2006), Yoshida and Ishii (2006).

[3]See, for instance, Ghahramani (2001) and Arcidiacono and Miller (2006).

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

1 Two-armed bandit "reversal learning" experiment

The learning experiments considered in this paper are adapted from Hampton, Bossaerts, and 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 good (bad) arm yields rewards

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

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[5]

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

.

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

Remark 1 ***(reversal learning):*** This bandit problem with reversal learning differs in important ways from the "standard" multi-armed bandit problem (cf. Gittins and Jones (1974), Banks and Sundarum (1992)), in which the states of the bandits are fixed over all periods and the bandits are "independent" in that a reward from one bandit is uninformative about the state of another bandit. The optimal Bayesian decision rule in the standard model features exploration (or "experimentation"), which recommends

[5] This aspect of our model differs from Hampton, Bossaerts, and 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.

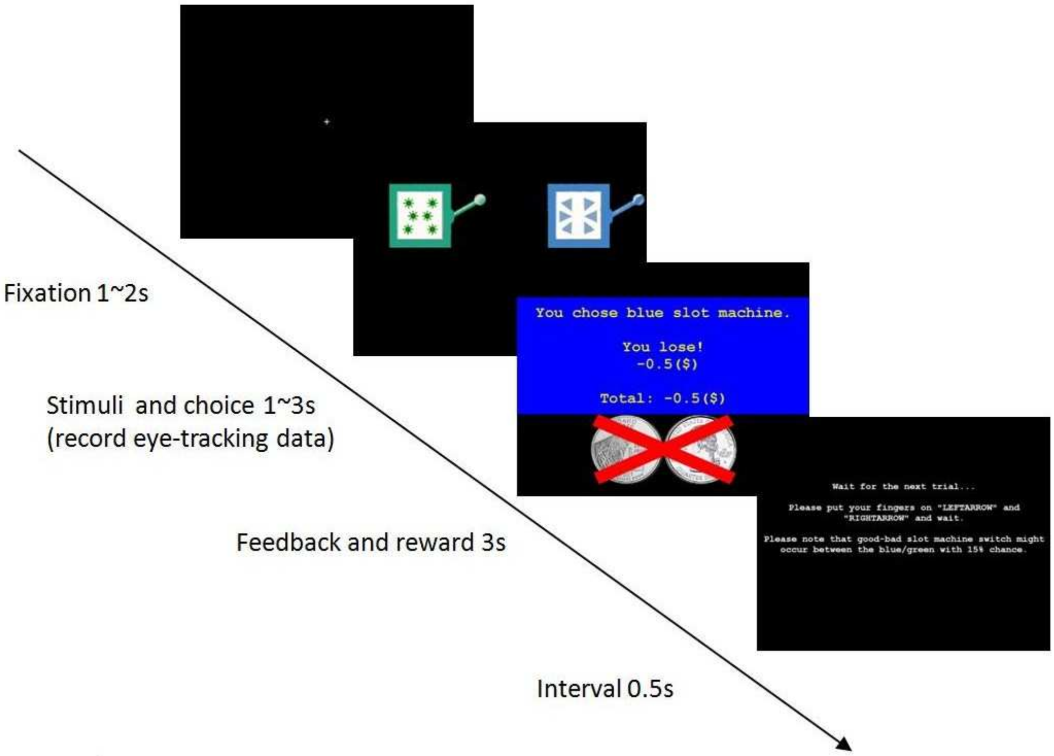


Figure 1: Timeline of a trial

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

sacrificing current rewards to achieve longer-term payoffs. In the setting considered in this paper, however, the bandits are negatively correlated, so that positive information about one slot machine implies negative information about the other. This eliminates most of the incentives for subjects to experiment.

1.1 Data

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,[6] 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 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.[7]

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 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.[8]

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

[6]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.

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

[8]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).

Table 1: Summary statistics for experimental data

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

	1(green)	2(blue)
Y	2108	2092

	1 ($0.50)	2 (-$0.50)
R	2398	1802

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

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

Z (after two-value discretization)^A	
1(green, $Z_p < 0$)	2(blue, $Z_p \geq 0$)
2032	2168

Z (after three-value discretization)^A		
1(green)	2(not sure)	3(blue)
1887	540	1773

^{A}: for more details on discretizing Z, see the appendix, section B

obtain if the subjects were choosing completely randomly.[9] Hence, subjects appear to be "trying", which motivates our analysis of their learning rules.

1.1.1 Remarks on eye-tracking measure

Because eye-tracking is still a relatively novel tool in economics, we present some discussion here. Recently, eye-tracking has been employed to assess subjects' thinking processes in various decision environments: to determine how subjects detect truth-telling

[9]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.

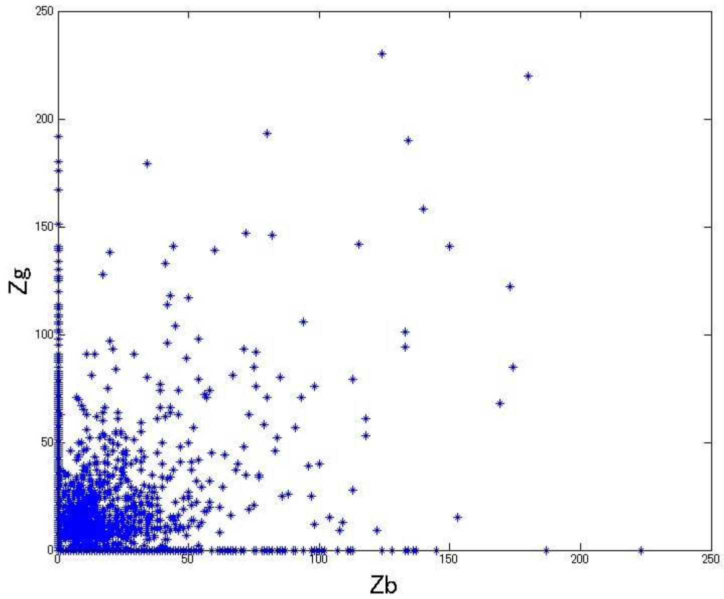


Figure 2: Scatter plot of Z_b (fixation on blue) and Z_g(fixation on green)
Both Z_b and Z_g are reported in 2×10^{-2} seconds.

or deception in sender-receiver games (Wang, Spezio, and Camerer (forthcoming)); how consumers evaluate comparatively a huge number of commodities, as in a supermarket setting (Reutskaja, Nagel, Camerer, and Rangel (forthcoming)); and the relationship between visual attention (as measured by eye-fixations) and valuation of commodities in choice tasks (cf. Krajbich, Armel, and Rangel (2007), Armel and Rangel (2008), Armel, Beaumel, and Rangel (2008), Rangel (2008)). Specifically, Armel and Rangel (2008) construct a plausible behavioral-neuroscientific model of value computation through visual attentions which successfully explains the observed relationship between fixation times and subjects' valuations in their experiments.[10]

In this paper, we use subjects' fixation durations as noisy measures of their beliefs (or valuations) for each slot machine. The raw eye-movement measure, $Z_{p,t}$, is defined as,

$$Z_{p,t} = (Z_{b,t} - Z_{g,t})/RT_t; \tag{2}$$

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

[10]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. Zhang, Wedel, and Pieters (2009)).

the subject's choice.[11] Thus, $Z_{p,t}$ measures how much longer a subject looks at the blue slot machine than the green one during the t-th trial, with a larger (smaller) value of $Z_{p,t}$ implying longer fixation time at the blue (green) slot machine. Summary statistics on this measure are given in the bottom panels of Table 1. Particularly, panel 4 shows that the correlation between Y_t (which =2(1) if blue(green) is chosen) and $Z_{p,t}$ is 0.7647, which suggests that in this choice setting, a longer fixation duration at an alternative implies a larger probability of choosing it. This also provides some justification for our use of eye movements as noisy measurements of subjects' beliefs, which affect their choices.

Figure 2 contains the scatter plot of $Z_{b,t}$ versus $Z_{g,t}$. The symmetric distribution around the 45-degree line in Figure 2 indicates that subjects are not intrinsically biased toward a certain color: the existing literature has reported that human subjects exhibit a "right side bias", tending to gaze towards the right side more frequently. However, our experimental data contains no significant evidence of such a bias. In our empirical work, we will discretize the eye-movement measure Z_p; to avoid confusion, in the following we use Z_p to denote the undiscretized eye-movement measure, and Z the discretized measure, which we describe below.

2 Econometric model

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

Assumption 2 *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).*

Assumption 3 *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 $P(X_t^*|X_{t-1}^*, R_{t-1}, Y_{t-1})$.*

[11]Furthermore, in order to control for subject-specific heterogeneity, we normalize $Z_{p,t}$ across subjects by dividing by the subject-specific standard deviation of $Z_{p,t}$, across all rounds for each subject.

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

Assumption 4 *The auxiliary measure Z_t is a noisy measure of beliefs X_t^*, with the measurement probabilities $P(Z_t|X_t^*)$. We assume that:*
(i) For all t, the $K \times K$ matrix $\mathbf{G}_{Z_t|Z_{t-1}}$, with $(i,j)-th$ entry equal to $Pr(Z_t = i|Z_{t-1} = j)$, is invertible.
(ii) $E[Z_t|X_t^*]$ is increasing in X_t^*.

The invertibility assumption 4(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 4(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.[12] The large correlation of 0.76 between Z_t and Y_t (as reported in Table 1 above) provides some indirect evidence favoring this monotonicity assumption.

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

Assumption 5 *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, blocks, and trials t.*

Remark 6 ***(stationary in learning models):*** An important benefit of considering a "probabilistic reversal" model (in which the identity of the "good" slot machine changes stochastically across trials) rather than the simpler standard multi-armed bandit model (in which the identity of the "good" arm is fixed across all trials) is that in the latter case, the subject's uncertainty regarding the identity of the "good" arm is decreasing across trials, so that learning rule must also condition on some measure of the subject's uncertainty (such as the number of times a particular arm has been pulled before a given trial) in order to satisfy the stationarity Assumption 4.[13] In a probabilistic reversal

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

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

setting, however, a subject's uncertainty does not decrease across trials. This is an attractive feature because, in our nonparametric estimation approach, conditioning on additional variables decreases the precision of the estimates.

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

2.1 Nonparametric identification

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

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

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:

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

Similarly to the above, we can derive that

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

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

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

From Assumption 4(i), we combine the two previous matrix equalities to obtain

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

This is an eigenvalue decomposition of the matrix $\mathbf{A}_{y,Z_t|Z_{t-1}} \mathbf{G}_{Z_t|Z_{t-1}}^{-1}$, which can be computed from the observed data sequence $\{Y_t, Z_t\}$.[14] This shows that from the observed

[14]Note that, from Eq. (5), the invertibility of $\mathbf{G}$ (which is Assumption 4(i)) implies the invertibility of $\mathbf{B}$.

data, we can identify the matrices $\mathbf{B}_{Z_t|X_t^*}$ and $\mathbf{D}_{y|X_t^*}$, which are the matrices with entries equal to (respectively) the measurement probabilities $P(Z_t|X_t^*)$ and choice probabilities $P(Y_t|X_t^*)$.

In order for this identification argument to be valid, the eigendecomposition in Eq. (6) must be unique. This requires the eigenvalues in this decomposition (corresponding to choice probabilities $P(y|X_t^*)$) to be distinctive; that is, $P(y|X_t^*)$ should vary in X_t^*. This is ensured by Assumption 2. Furthermore, even if the eigendecomposition is unique, the representation in Eq. (6) is invariant to the ordering (or permutation) and scalar normalization of eigenvectors. Assumption 4(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 $\mathbf{P}(\mathbf{X}_{t+1}^*|\mathbf{X}_t^*, \mathbf{R}_t, \mathbf{Y}_t)$. Again, start with a factorization

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

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

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

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

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

Assumption 5 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

$$
\mathbf{L}_{X_{t+1}^*, X_t^*, y, r} = \mathbf{B}_{Z_{t+1}|X_{t+1}^*}^{-1} \mathbf{H}_{Z_{t+1},y,r,Z_t} [\mathbf{B}'_{Z_t|X_t^*}]^{-1}. \tag{8}
$$

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

3 Estimation

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

Our estimation procedure 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. (6) and (8).

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 the estimates do go outside this range. In these cases, we obtained the estimates by a least-squares fitting procedure, where we minimized the elementwise sum-of-squares corresponding to Eqs. (6) and (8), and explicitly restricted each element of the matrices to lie $\in [0, 1]$. This was not a frequent recourse; only a handful of the estimates reported below needed to be restricted in this manner.

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

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

3.1 Estimation results

3.1.1 Two-value estimates

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

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

Remark 7 ***(What do the beliefs $\{X_t^*\}$ mean?)*** As we discussed earlier in Remark 1, in the standard multi-armed bandit model, subjects' choices of which arm to pull depends on the dynamic allocation, or "Gittins" index, which depends not only on current beliefs about which arm yields a higher return, but also on the informational value in pulling an arm which may not be currently optimal, but which may yield information useful in future decisions. However, in our reversal learning setting, because the returns in the two arms are negatively correlated, this informational value term is nonexistent. Therefore, in the context of such a model, we can quite confidently interpret the unobserved variables X_t^*, which completely determine subjects' choices in our learning model, as a measurement of subjects' current beliefs regarding which arm is currently the "good" one. Thus another benefit of a reversal learning model is the unambiguity in interpreting the unobserved "beliefs" X_t^* in this setting.

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

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

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

$P(Y_t|X_t^*)$

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

$P(Z_t|X_t^*)$

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

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

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

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

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

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

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

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

$P(Y_t|X_t^*)$

X_t^*	1(green)	2(blue)
$Y_t = 1$ (green)	0.9756 (0.0115)	0.0573 (0.0165)
2 (blue)	0.0244	0.9427

$P(Z_t|X_t^*)$

X_t^*	1(green)	2(blue)
$Z_t = 1$ (green)	0.9093 (0.0156)	0.0888 (0.0116)
2 (blue)	0.0907	0.9112

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

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

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

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

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

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

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

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

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

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

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

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

At face value, this suggests that subjects do not update their beliefs properly following "exploratory" (ie. contrary to belief) actions. However, as we will see now, these puzzling results are less apparent when we allow beliefs to take three distinct values.

3.1.2 Three-value estimates

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

Table 4 contains the estimates of the choice and measurement probabilities.[15] The

[15]We also considered a robustness check against the possibility that subjects' fixations immediately

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

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

Choice probabilities:

$P(Y_t|X_t^*)$

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

$P(Z_t|X_t^*)$

X_t^*	1(green)	2(not sure)	3(blue)
$Z_t = 1$ (green)	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 (blue)	0.0546 (0.0581)	0.1499 (0.1206)	0.8421 (0.0529)

first and last columns of the panels in this table indicate that choices and eyes movements are closely aligned with beliefs, when beliefs are sufficiently strong (ie. are equal to either $X^* = 1$ or $X^* = 3$). Specifically, in these results, the "exploration probability" is smaller than in the two-value results, being equal to 1.3% when $X_t^* = 1$, and only 0.64% when $X_t^* = 3$. As we discussed in Remark 1 above, such small probabilities can be consistent with optimal behavior, in our reversal learning environment, where subjects have little incentive to experiment.

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

The learning rule estimates are presented in Table 5. The results are similar to the two-value results, but most of the problems from those results disappear when we

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 5: Three-value estimates: Specification where X_t^* and Z_t take three values

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

Learning Rule updating probabilities:

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

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

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

X_t^*	1(green)	2 (not sure)	3(blue)
$X_{t+1}^* = 1$ (green)	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 (blue)	0.1111 (0.0340)	0.0678 (0.0485)	0.0000 (0.1876)

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

X_t^*	3(blue)	2 (not sure)	1(green)
$X_{t+1}^* = 3$ (blue)	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 (green)	0.4166 (0.0874)	0.5607 (0.0968)	0.6792 (0.1881)

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

X_t^*	3(blue)	2 (not sure)	1(green)
$X_{t+1}^* = 3$ (blue)	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 (green)	0.1155 (0.0499)	0.0279 (0.0373)	0.0116 (0.0679)

allow beliefs to take three values. The left columns show how beliefs are updated when "exploitative" choices (ie. choices made in accordance with beliefs) are taken. When current beliefs indicate "green" ($X_1^* = 1$) and green is chosen ($Y_t = 1$), beliefs are quite responsive to the reward: if $R_t = 1$ (the low reward), then beliefs stay at green with probability 57%, but if $R_t = 2$ (high reward), then this probability is much higher, at 89%. On the other hand, even after positive (ie. high reward) exploitative choices, beliefs may still update towards "blue" ($X_{t+1}^* = 3$) with an 11% chance, rather than staying at the intermediate level $X_{t+1}^* = 2$. This non-smooth "extremal" updating is a distinctive feature of our learning rule estimates, and is consistent with optimal belief-updating in a probabilistic reversal context: even if the subject were completely sure that "green" after a high reward, she still must consider the possibility that the good state could change to "blue" by the next trial, due to the stochastic evolution of the state process.

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

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

4 Comparing nonparametric vs. standard learning models

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

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

A noteworthy feature is that the histograms for the eye-tracking measure Z_p and the TD-learning valuations look similar: both are trimodal. The Bayesian posterior mean measure, on the other hand, is unimodel. As we will see later, this implies that beliefs from the nonparametric model will be closer to the RL model, than the Bayesian model. Moreover, we will also see that the Bayesian learning model tends to predict "smoother" choice behavior than what we observe in the data, while the beliefs from the nonparametric model are "jumpy" in comparison.

Overall summary statistics In Table 6, we present some summary statistics for the implied beliefs from our nonparametric learning model (denoted X_t^*), vs. the Bayesian beliefs B^* and the valuations V^* in the RL learning model. For simplicity, we will abuse terminology somewhat and refer in what follows to X^*, V^*, and B^* as the "beliefs" implied by, respectively, our nonparametric model, the RL model, and the Bayesian model. This table contains eight panels.

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

Panel 3 contains the pairwise correlation among (X^*, V^*, B^*), the beliefs from the three models. The correlation between X^* and V^* (0.59) exceeds that between X^* and B^* (0.53). This shows that the nonparametric beliefs X^* are, stochastically, more similar to the RL beliefs V^* than to the Bayesian beliefs B^*. This finding confirms the evidence from the histograms, as described above. The Bayesian model is the most restrictive one,

Figure 3: Histogram of Z_p

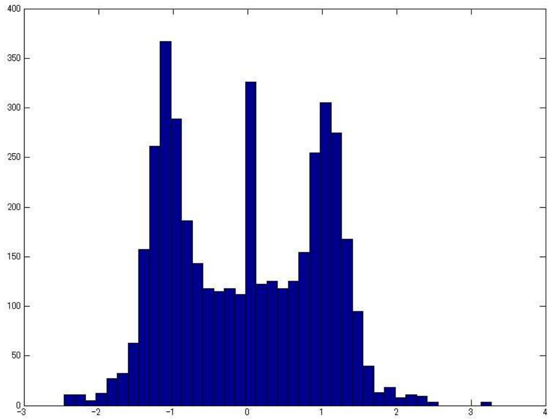


Figure 4: Histogram of Bayesian Belief B^*

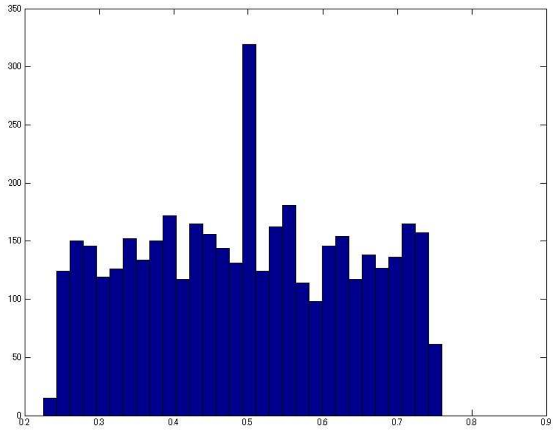


Figure 5: Histogram of $V^* = V_b - V_g$ in RL

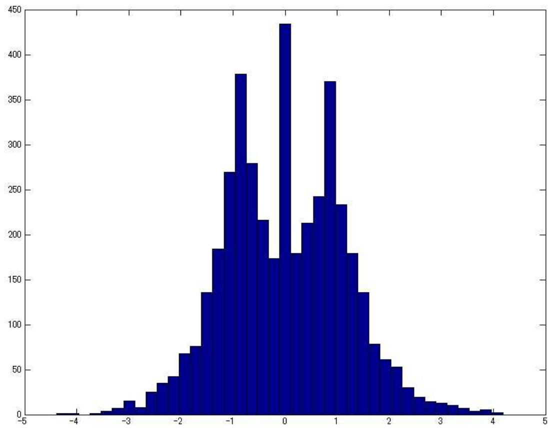


Table 6: Summary statistics for the three models

Panel 1:

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

Panel 2:

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

Panel 3: Correlations in the three models

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

*: bootstrapped standard error in parentheses

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

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

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

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

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

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

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

Nonparametric	402
Reinforcement Learning	455
Bayesian	543

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

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

and imposes the highest degree of rationality on subjects, which may explain its inferior fit, relative to the RL model, to our nonparametric learning rule.

However, the correlations between our nonparametric beliefs X^* and B^* and V^* are markedly lower than that between B^* and V^* (which is 0.82). This indicates that, informationally, the beliefs from the Bayesian and RL models are very similar.

The next panel shows that the correlation of X^* with the observed choices Y is higher (0.7552) than the correlation of choices with the beliefs from the other models. This superior performance of the nonparametric beliefs in predicting subjects' choices is not too surprising, since the beliefs are estimated from the data, whereas the other two models are only calibrated to the data. The next two panels break down the correlation between the observed choices and the difference measures of beliefs, for subsamples of the data. Panel 5 only considers subjects' choices when the implied beliefs are strong (in the sense of taking extreme values). For the nonparametric model, we omitted observations when X^* was estimated to be "not sure", while for the other two models, we omitted observations when beliefs lay between the 1/3 and 2/3 quantile. The results show that when beliefs are strong, the nonparametric model continues to predict choices better than the Bayesian and RL models. Panel 6 shows that predicted choice behavior is more accurate (using all three models) during the last ten rounds of each subject's data, and less accurate during the first five rounds. This supports the notion that subjects behaved more haphazardly at the beginning of the experiments.[16]

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

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

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

require Z to compute, exhibit a smaller correlation with Z.

A closer look at individual blocks To look more closely at the differences between the three learning models, we plot, in Figures 6-9, the actual choices, as well as subjects' beliefs regarding which slot is better, from the three learning models, for four representative subject-blocks of choices. The actual choices are plotted in crosses (+'s), with higher crosses (at 0.25) signifying "blue" and lower crosses (at -0.25) signifying "green". The subject's beliefs from the three models, all recentered and rescaled around zero, are plotted; X_t^* as a solid line, B_t^* dotted, and V_t^* dashed.[17]

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

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

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

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

Figure 6: Subject 6, block 4

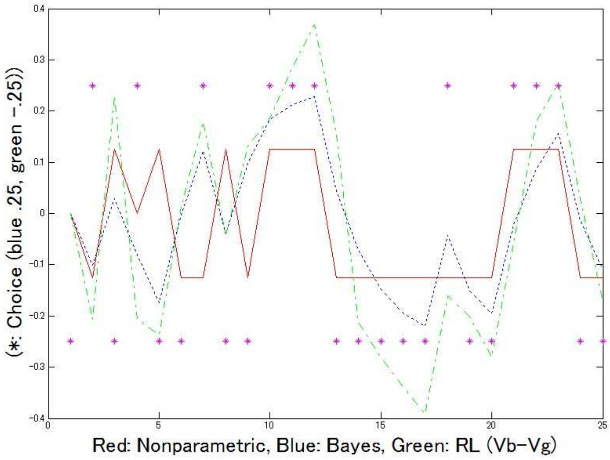


Figure 7: Subject 4, block 6

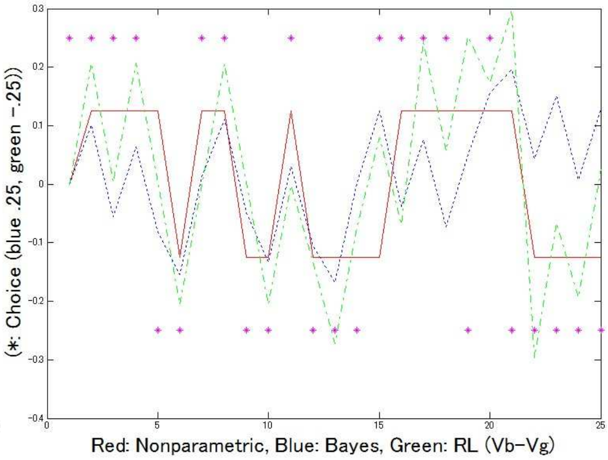


Figure 8: Subject 5, block 8

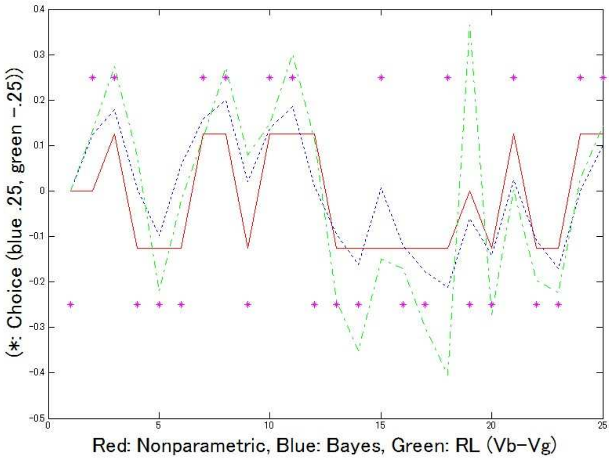
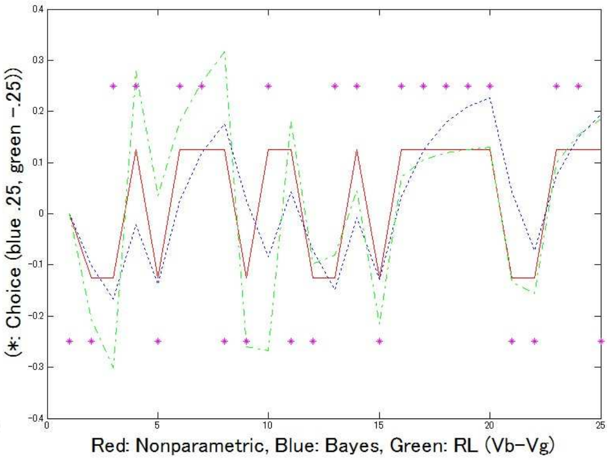


Figure 9: Subject 1, block 3



learning rule represent another important qualitative difference relative to the standard learning models.

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, non-smooth updating following positive "exploitative" choices. A comparison of the nonparametric learning rules with "standard" learning models shows that our estimates are closer to the reinforcement learning model, than to a Bayesian model. Altogether, our analysis points out some deficiencies in the Bayesian model as a descriptive model, thus echoing previous findings in both the experimental and finance literatures.

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

References

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Miller, R. (1984): "Job Matching and Occupational Choice," *Journal of Political Economy*, 92, 1086–1120.

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

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

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

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

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

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

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

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

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

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

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

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.

Appendix A Appendix: Details on computation of beliefs in the nonparametric, Bayesian, and RL 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.

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

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

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

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

$$\log L(\{X_t^*, Z_t\}|\{Y_t, R_t\}) = \sum_{t=1}^{24} \log \left[P(Z_t|X_t^*)P(X_{t+1}^*|X_t^*, R_t, Y_t)\right] + \log(P(Z_{25}|X_{25}^*)). \tag{9}$$

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).

Note that, 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 (9) above. This was because, given our estimates that $P(Y_t = 1|X_t^* = 1) \approx P(Y_t = 2|X_t^* = 3) \approx 1$ in Table 4, maximizing with respect to these choice probabilities would leads to estimates of beliefs $\{X^*\}$ which closely coincide with observed choices; we wished to avoid such an artificially good "fit" between the beliefs and observed choices.

For robustness, however, we also estimated the beliefs $\{X^*\}$ under two alternative scenarios: (i) treating the choice sequence $\{Y_t\}$ as endogenous, and hence including the choice probabilities $P(Y_t|X_t^*)$ in the likelihood function; (ii) treating both $\{Y_t, Z_t\}$ as exogenous, and hence omitting both the choice probabilities $P(Y_t|X_t^*)$ and the measurement probabilities $P(Z_t|X_t^*)$ from the likelihood function. Not surprisingly, the correlation between choices and beliefs $\text{Corr}(Y_t, X_t^*) = 0.99$ under (i), while under (ii) the correlation falls to 0.56. However, in both of these alternative specifications, we still find that $\text{Corr}(X_t^*, V_t^*) > \text{Corr}(X_t^*, B_t^*)$ – that is, the nonparametric beliefs are "closer" to the RL model than the Bayesian model. Thus this finding appears robust across a number of different approaches to recovering the nonparametric beliefs $\{X_t^*\}$.

A.2 Bayesian Learning Model

A Bayesian learner uses Bayes rule to update her beliefs. Let B_t^* denote the prior probability that the blue slot machine is the good one at the start of the trial t. After her choice Y_t, she observes reward R_t, and updates her belief that the blue slot machine is good to B'^*_t; by Bayes' rule, this updated probability is:

$$B'^*_t = \frac{P(R_t|Y_t, S_t = 1) \cdot B_t^*}{P(R_t|Y_t, S_t = 1) \cdot B_t^* + P(R_t|Y_t, S_t = 2) \cdot (1 - B_t^*)} \tag{10}$$

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

$$B_{t+1}^* = P(S_{t+1} = 1|S_t = 1) \cdot B'^*_t + P(S_{t+1} = 1|S_t = 2) \cdot (1 - B'^*_t). \tag{11}$$

In this way, given the initial beliefs $B_1 = 0.5$, we can use Eqs. (10) and (11) to compute the sequence of Bayesian beliefs, $\{B_t^*\}$, corresponding to the observed sequences of choices and rewards $\{Y_t, R_t\}$. The corresponding choice rule from the Bayesian model would be to choose "blue" at trial t iff $B_t^* \geq 0.5$.

A.3 Reinforcement Learning Model

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

TD-Learning model is defined as:

$$V_{Y_t}^{t+1} \longleftarrow V_{Y_t}^t + \alpha \delta_t. \tag{12}$$

where Y_t denotes the choice taken in trial t, α denotes the learning rate, and δ_t denotes the "prediction error" for trial t (defined below). The prediction error δ_t is equal to

$$\delta_t = (R_t + \gamma E[V_{Y_{t+1}}^t | t]) - V_{Y_t}^t \tag{13}$$

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

The variant of TD-Learning (SARSA, short for "state-action-reward-state-action") used here (Sutton and Barto (1998), p. 149) computes the expected value function $E[V_{Y_{t+1}} | t]$ using the current choice probabilities of choosing the future action Y_{t+1} (which is unknown at trial t). P_c^t, the current probability of choosing action c, is assumed to take the conventional "softmax" (ie. logit) form with the inverse temperature parameter β:

$$P_c^t = e^{\beta V_c^t} / \left[\sum_{c'} e^{\beta V_{c'}^t} \right] \tag{14}$$

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

$$E[V_{Y_{t+1}} | t] = \sum_{c' \in (b,g)} P_{c'}^t V_{c'}^t. \tag{15}$$

We estimated the parameters β and α using maximum likelihood. For greater model flexibility, we allowed the parameter α to differ following positive vs. negative rewards. (We fixed the discount rate $\gamma = 0.9$.) The estimates we obtained from the data were:

$$\begin{aligned} &\beta = 0.7584 \\ &\alpha \text{ for positive reward } (R_t = 2) = 1.6531 \\ &\alpha \text{ for negative reward } (R_t = 1) = 1.0552. \end{aligned} \tag{16}$$

We plug in these values into Eqs. (12), (13), (14) and (15) to derive a sequence of valuations $\left\{ V_t^* \equiv V_b^t - V_g^t \right\}$. The choice function (Eq. (14)) can be rewritten as a function of the difference V_t^*; i.e. the choice probability for the blue slot machine is,

$$P_b^t = \frac{e^{\beta(V_b^t - V_g^t)}}{1 + e^{\beta(V_b^t - V_g^t)}} = \frac{e^{\beta V_t^*}}{1 + e^{\beta V_t^*}} \tag{17}$$

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.

Appendix B Appendix: Details on discretization

In this section, we present additional discussion on the discretization of the eye-movement measure, and some evidence that a three-valued discretization (which we used in our preferred empirical specifications) is sufficient to capture most of the variation in this measure. Let $Z_{p,t}$ denote the continuous-valued eye-tracking measure, and Z_t the discretized version, both for trial t. For the two-value discretization, we discretize as follows:

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

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

$$Z_t = \begin{cases} 1 & \text{if } Z_{p,t} < -\sigma_z \\ 2 & \text{if } -\sigma_z \leq Z_{p,t} \leq \sigma_z \\ 3 & \text{if } \sigma_z < Z_{p,t} \end{cases} \tag{18}$$

where σ_z denotes a constant used to discretize $Z_{p,t}$. 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 σ_z from 0.05 to around 0.40, suggesting that the model is robust for different classifications. Table 1 contains the summary statistics for both the two- and three-value discretizations. Table 7 shows the sample frequencies of the discretized measure Z_t for three different values of σ_z.

Figure 3 is the histogram of $Z_{p,t}$, which is apparently trimodel, with peaks at -1, 0 and 1, which suggest that a three-value discretization of Z_p indeed captures most of its variation. Moreover, Table 7 shows the correlations between Y and Z_p, broken up into the three ranges of Z_p corresponding to the three discretized values $Z \in \{1, 2, 3\}$, and also for three different values of the σ_z parameter. Although the correlation in the whole sample is 0.7647, the correlations within each of the three ranges of Z_p 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.

Table 7: Correlations between (Y, Z_p) in different subsamples

	Size	Corr(Y, Z_p)
Full sample	4200	0.7647
σ_z = 0.20 (baseline):		
$Z = 1$ (green)	1887	0.2845
2 (not sure)	540	0.2156
3 (blue)	1773	0.1706
σ_z=0.05:		
$Z = 1$ (green)	2015	0.3223
2 (not sure)	255	-0.0599
3 (blue)	1930	0.2346
σ_z=0.40:		
$Z = 1$ (green)	1725	0.1462
2 (not sure)	869	0.2777
3 (blue)	1606	0.0991

Note: Z_p refers to the undiscretized eye-movement measure, as defined in Eq. (2), and Z refers to the discretized version, as defined in Eq. (18).