Games and Economic Behavior 81 (2013) 215–231

Contents lists available at SciVerse ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Nonparametric learning rules from bandit experiments: The eyes have it! ☆

Yingyao Hu [a,*], Yutaka Kayaba [b], Matthew Shum [b]

[a] Johns Hopkins University, Dept. of Economics, 3400 North Charles Street, Baltimore, MD 21224, United States
[b] Caltech, Division of Humanities and Social Sciences, 1200 East California Blvd., Pasadena, CA 91125, United States

ARTICLE INFO

Article history:
Received 22 February 2012
Available online 30 May 2013

JEL classification:
D83
C91
C14

Keywords:
Learning
Belief dynamics
Experiments
Eye tracking
Bayesian vs. non-Bayesian learning

ABSTRACT

How do people learn? We assess, in a model-free manner, subjects' belief dynamics in a two-armed bandit learning experiment. A novel feature of our approach is to supplement the choice and reward data with subjects' *eye movements* during the experiment to pin down estimates of subjects' beliefs. Estimates show that 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 average earnings by subjects in this experiment) than what would be obtained from a fully-rational Bayesian learning model, but comparable to the profits from alternative non-Bayesian learning models, including reinforcement learning and a simple "win-stay" choice heuristic.

How do individuals learn from past experience in dynamic choice environments? A growing literature has documented, using both experimental and field data, that the benchmark fully-rational Bayesian learning model appears deficient at characterizing actual decision-making in real-world settings.[1] Other papers have demonstrated that observed choices in strategic settings with asymmetric information are typically not consistent with subjects' having Bayesian (equilibrium) beliefs regarding the private information of their rivals.[2]

Given the lack of consensus in the literature about what the actual learning rules used by agents in real-world decision environments are, there is a need for these rules to be estimated in a manner flexible enough to accommodate alternative models of learning. In this paper, we propose a new approach for assessing agents' belief dynamics. In an experimental setting, we utilize data on subjects' *eye movements* during the experiment to aid our inference regarding the learning (or belief-updating) rules used by subjects in their decision-making process. Taking advantage of recent developments in the

☆ 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, Mark Dean, Cary Frydman, Ian Krajbich, Pietro Ortoleva, Joseph Tao-yi Wang and participants in presentations at U. Arizona, Caltech, UCLA, U. Washington and Choice Symposium 2010 (Key Largo) for comments and suggestions. Kayaba thanks the Nakajima Foundation for the financial support.

* Corresponding author.
E-mail addresses: yhu@jhu.edu (Y. Hu), ykayaba@caltech.edu (Y. Kayaba), mshum@caltech.edu (M. Shum).

[1] An incomplete list of papers which consider these questions includes El-Gamal and Grether (1995), Nyarko and Schotter (2002), Charness and Levin (2005), Kuhnen and Knutson (2008), and Payzan-LeNestour and Bossaerts (2011).

[2] For instance, Bajari and Hortaçsu (2005), Goldfarb and Xiao (2011), Ho and Su (2010), Crawford and Iriberri (2007), Gillen (2011), Brown et al. (2012), Brocas et al. (2009).

0899-8256/$ – see front matter
http://dx.doi.org/10.1016/j.geb.2013.05.003

econometrics of dynamic measurement error models, we use the experimental subjects' choice and eye-tracking data to estimate subjects' decision rules and learning rules, without imposing a priori functional forms on these functions.

Eye-movement data are an example of novel non-choice variables, the measurement and analysis of which has constituted an important strand in experimental economics. Caplin and Dean (2008) broadly call these auxiliary measures "neuroeconomic" data, in the sense of data other than the usual choice and rewards data gathered from typical experiments. Besides eye tracking (Krajbich et al., 2010; Wang et al., 2010; Knoepfle et al., 2009), other examples of such data include measurements of brain activity (Boorman et al., 2009; Hsu et al., 2005; Preuschoff et al., 2006), pupil dilation (Preuschoff et al., 2011), skin conductance response (Sokol-Hessner et al., 2009) and mouse tracking (Camerer et al., 1993; Johnson et al., 2002; Costa-Gomes et al., 2001; Costa-Gomes and Crawford, 2006; Brocas et al., 2009; Gabaix et al., 2006).[3]

Our main results are as follows. First, our estimated learning rules do not correspond to any one of the existing learning models. Rather, we find that beliefs are reward-asymmetric, in that subjects are more reluctant to "update down" following unsuccessful (low-reward) choices, than "update up" following successful (high-reward) choices. Moreover, from a payoff perspective, these learning rules are suboptimal relative to the fully-rational Bayesian benchmark. Specifically, using the estimated learning rules, subjects' payoffs are, at the median, \$4 (or about two cents per choice) lower than under the Bayesian benchmark; this difference represents about 25% of average earnings of subjects in this experiment (not including the fixed show-up fee). However, subjects' payoffs under the estimated choice and learning rules are comparable to the profits from alternative non-Bayesian learning models, including reinforcement learning.

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 a model of subjects' choices in the bandit model, and discuss estimation. In Section 3, we present our 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 4 concludes.

1. Two-armed bandit "reversal learning" experiment

Our experiments are adapted from a "reversal learning" model used in Hampton et al. (2006). This is a simplified two-armed bandit model which is ideal for studying subjects' learning and belief updating while abstracting away from strategic concerns, which are absent in this model.[4] 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 reward R_t that the subject receives in trial t depends on the action taken, as well as (stochastically) on the current state: if the good arm is chosen, then the reward is:

$$R_t = \begin{cases} +\$0.50 & \text{with prob. } 70\%, \\ -\$0.50 & \text{with prob. } 30\%. \end{cases} \tag{1}$$

If the bad arm is chosen, the reward is generated as:

$$R_t = \begin{cases} +\$0.50 & \text{with prob. } 40\%, \\ -\$0.50 & \text{with prob. } 60\%. \end{cases} \tag{2}$$

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[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

. (3)

[3] Our paper also joins other papers which have considered the statistical or econometric analysis of neuroeconomic data. These papers include: Costa-Gomes et al. (2001), Costa-Gomes and Crawford (2006), Hsu et al. (2005, 2009), Caplin et al. (2010).

[4] Bandit problems are frequently used to model various economic decision-problems focusing on explore-exploitation trade-off and belief updating, such as pricing in a market with unknown demand (Rothschild, 1974), R&D choice (Weitzman, 1979), labor markets (Jovanovic, 1979) and financial investment (Bergemann and Hege, 1998), as well as experimental studies (Banks et al., 1997; Meyer and Shi, 1995; Gans et al., 2007; Anderson, 2012). The probabilistic reversal model focuses on learning about regime switches in its simplest form, which makes it particularly relevant for financial applications (e.g. how agents change investment strategies depending on news on market fundamentals).

[5] This aspect of our model differs from Hampton et al. (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 require alternative identification arguments than the one considered in this paper.

Because S_t is not observed by subjects, and is serially-correlated over time, subjects have an opportunity 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.

1.1. Benchmark: Fully-rational decision-making in reversal learning model

For comparison with our empirical model and results, which are presented below, we discuss here how an optimal dynamic decision-maker would behave in a probabilistic reversal framework. First we introduce some notation and describe the information structure and how Bayesian updating would proceed in the reversal learning context. Let Q denote the 2×2 Markov transition matrix for the state S_t, corresponding to Eq. (3).

Let B_t^* denote the probability (given by Bayes' rule) that a subject places on "blue" being the good arm in period t, conditional on the whole experimental history up to then. Specifically, B_t^* denotes the belief that $S_t = 2$, at the beginning of period t, while $\tilde{B}_t^*$ denotes the updated belief that $S_t = 2$, 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 Bayes' rule:

$$\tilde{B}_t^* = P\left(S_t = 2 \middle| B_t^*, R_t, Y_t\right) = \frac{B_t^* \cdot P(R_t|S_t = 2, Y_t)}{(1 - B_t^*) \cdot P(R_t|S_t = 1, Y_t) + B_t^* \cdot P(R_t|S_t = 2, Y_t)}.$$

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

$$\begin{bmatrix} 1 - B_{t+1}^* \\ B_{t+1}^* \end{bmatrix} = Q \cdot \begin{bmatrix} 1 - \tilde{B}_t^* \\ \tilde{B}_t^* \end{bmatrix} = Q \cdot \begin{bmatrix} 1 - P(S_t = 2|B_t^*, R_t, Y_t) \\ P(S_t = 2|B_t^*, R_t, Y_t) \end{bmatrix}. \tag{4}$$

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.)

We evaluated the fully-rational decision rules – the mapping from period t beliefs B_t^* to a period t choice – in this dynamic Bayesian learning model by computer simulation. The optimal decision rules for the model have two important features. First, the decision rules are identical across all the periods, indicating that they are *stationary*. Second, the fully-rational decision rule takes a simple form: in each period, it prescribes that subjects choose the blue arm whenever the current belief B_t^* that the blue arm is "good" exceeds 50%. This is a *myopic* decision rule.

These optimal decision rules for the reversal learning model differ in important ways from optimal decision-making in the standard multi-armed bandit (MAB) problem (cf. Gittins and Jones, 1974), 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 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.[6] In our reversal learning setting, however, the states of the bandits are negatively related, so that positive information about one arm implies negative information about the other. This negative correlation between the arms eliminates the incentives for subjects to experiment.

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,[7] 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.[8]

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. Fig. 1 contains 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 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.[9]

[6] See Miller (1984), Erdem and Keane (1996), Ackerberg (2003), Crawford and Shum (2005), and Chan and Hamilton (2006) for empirical studies of learning and experimentation in a dynamic choice context.

[7] 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.

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

[9] 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).

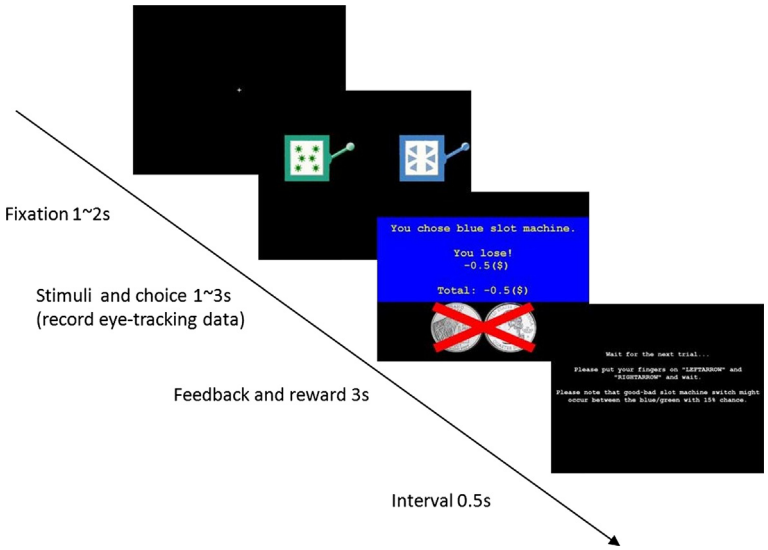

Fig. 1. Timeline of a trial. After subjects fix their gaze on the cross (top screen), two slot machines are presented (second screen). Subjects' eye movements are recorded by the eye-tracking machine here. Subjects choose by pressing the left (right) arrow key to indicate a choice of the left (right) 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%.

Table 1
Summary statistics for experimental data.

	1 (green)	2 (blue)
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[a]	−0.0309	0	1.3987	−1.4091
RT: reaction time (10^{-2} s)	88.22	59.3	212.2	36.8

	Y_t	Y_{t-1}
Correlation with $\tilde{Z}_t$	0.7642	0.3838

[a] Defined in Eq. (5).

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% ($\approx 2398/(2398+1802)$ with standard error 0.007637). This is slightly higher than, but significantly different from, 55%, which is the frequency which would obtain if the subjects were choosing completely randomly.[10] Hence, subjects appear to be "trying", which motivates our analysis of their learning rules. On the other hand, simulation of the fully-rational Bayesian decision rules (discussed above) shows that the success rate from using the fully-rational 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%.

Table 2 contains the empirical frequencies 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 a "reduced-form" decision rule for the subjects. The top row in that table contains the empirical frequencies, across all subjects, that the green arm is chosen, conditional on the different values of $(Y_t|Y_{t-1}, R_{t-1})$. 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, on average, subjects appear to replay successful strategies, corresponding to a "win-stay" rule-of-thumb.

[10] This is the marginal probability of a good reward, which equals $0.5(0.7+0.4)$ from Eqs. (1) and (2). The t-statistic for the null that subjects are choosing randomly equals 169.67, so that hypothesis is strongly rejected.

Table 2
"Reduced-form" decision rules: $P(Y_t = 1(\text{green})|Y_{t-1}, R_{t-1})$. Choice probabilities conditional on past choice Y_{t-1} and reward R_{t-1}.

$(Y_{t-1}, R_{t-1}) =$	(1, 1)	(1, 2)	(2, 1)	(2, 2)
Across all subjects:	0.5075	0.8652	0.5089	0.1189
	(0.0169)	(0.0094)	(0.1169)	(0.0090)
For each individual subject:				
Subject #1:	0.1799	0.5192	0.8128	0.364
	(0.0655)	(0.0684)	(0.0595)	(0.0603)
Subject #2:	0.1051	0.9820	0.9449	0
	(0.0498)	(0.0171)	(0.0381)	(0)
Subject #3:	0.7938	0.9859	0.3340	0
	(0.0591)	(0.0136)	(0.0871)	(0)
Subject #4:	0.3244	0.8796	0.6492	0.0610
	(0.0704)	(0.0514)	(0.0726)	(0.0283)
Subject #5:	0.0419	0.8796	0.6492	0.0610
	(0.0292)	(0.0236)	(0.0325)	(0.0461)
Subject #6:	0.2570	0.7498	0.8159	0.2021
	(0.0652)	(0.0592)	(0.0602)	(0.0532)
Subject #7:	0.5792	0.9242	0.4647	0.0796
	(0.0751)	(0.0371)	(0.0731)	(0.0379)
Subject #8:	0.8931	0.9803	0.1013	0.0165
	(0.0496)	(0.0186)	(0.0482)	(0.0163)
Subject #9:	0.6377	1.0000	0.2741	0
	(0.0831)	(0)	(0.0655)	(0)
Subject #10:	0.1986	0.9344	0.8037	0
	(0.0622)	(0.0352)	(0.0587)	(0)
Subject #11:	0.7859	1.0000	0.4306	0
	(0.0575)	(0)	(0.0870)	(0)
Subject #12:	0.5883	0.9262	0.3741	0.0131
	(0.0841)	(0.0406)	(0.0733)	(0.0129)
Subject #13:	0.6741	0.8907	0.1962	0.2085
	(0.0705)	(0.0462)	(0.0581)	(0.0539)
Subject #14:	0.4730	0.6147	0.5363	0.3842
	(0.0831)	(0.0653)	(0.0735)	(0.0664)
Subject #15:	0.6759	0.9789	0.3351	0
	(0.0761)	(0.0206)	(0.0714)	(0)
Subject #16:	0.4595	0.9135	0.5443	0.1953
	(0.0715)	(0.0316)	(0.0742)	(0.0666)
Subject #17:	0.6358	0.5202	0.5322	0.4644
	(0.0660)	(0.0706)	(0.0780)	(0.0748)
Subject #18:	0.6333	1.0000	0.2901	0
	(0.0834)	(0)	(0.0734)	(0)
Subject #19:	0.6144	0.8197	0.5808	0.2013
	(0.0702)	(0.0444)	(0.0806)	(0.0625)
Subject #20:	0.3699	0.5741	0.3699	0.3554
	(0.0858)	(0.0707)	(0.0665)	(0.0621)
Subject #21:	0.6990	0.9602	0.2934	0.0177
	(0.0658)	(0.0274)	(0.0693)	(0.0171)

Note: Standard errors (in parentheses) computed using 1000 block-bootstrap resamples.

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 (i.e. the first and third entries in this row). Thus, subjects tend to randomize after unsuccessful strategies. 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 empirical frequencies, 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, some subjects continue replaying an unsuccessful strategy: for example, 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.[11]

[11] 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 (i.e. yielding a low reward) may indeed be good in the next period.

1.3. Remarks on eye-tracking measure

Eye tracking, which involves the measurements of subjects' eye fixations during experiments, is a relatively new tool in economics, and we provide some background here. Recently, eye tracking has been employed to measure in various decision environments: to determine how subjects detect truth-telling or deception in sender–receiver games (Wang et al., 2010); how consumers evaluate comparatively a large number of commodities, as in a supermarket setting (Reutskaja et al., 2011); and the relationship between visual attention (as measured by eye fixations) and choices of commodities in choice tasks (cf. Krajbich et al., 2010; Armel and Rangel, 2008; Armel et al., 2008). The use of eye tracking followed upon the earlier use of "mouse tracking", which allowed experimenters a ways to track subjects' inferences and cognitive processes during an experiment.[12]

Our use of eye movements in this paper is predicated on an assumption that gaze is related to beliefs of expected rewards.[13] This is motivated by some recent results in behavioral neuroscience. Shimojo et al. (2003) studied this in binary "face 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 tend to be directed to the preferred face. Interestingly, the relationship between gaze direction and the chosen face weakens 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 Lauwereyns et al. (2002) and Kawagoe et al. (1998) (see the survey in Hikosaka et al., 2006), which shows that primates tend to direct their gaze at locations where rewards are available.[14]

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

$$\tilde{Z}_t = (Z_{b,t} - Z_{g,t})/RT_t; \tag{5}$$

that is, for trial t, $Z_{b(g),t}$ is the gaze duration at the blue (green) slot machine, and RT_t is the reaction time, i.e. the time between the onset of the trial after fixation, and the subject's choice. Furthermore, in order to control for subject-specific heterogeneity, we normalize $\tilde{Z}_t$ across subjects by dividing by the subject-specific standard deviation of $\tilde{Z}_t$, across all rounds for each subject. Thus, $\tilde{Z}_t$ measures how much longer a subject looks at the blue slot machine than the green one during the tth trial, with a larger (smaller) value of $\tilde{Z}_t$ implying longer gazes directed 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 machines. In our empirical work, we discretize the eye-movement measure; to avoid confusion, in the following we use $\tilde{Z}_t$ to denote the undiscretized eye-movement measure, and Z_t the discretized measure.[15]

2. Model

In this section, we introduce a model of dynamic decision-making in the two-armed bandit experiment described above, and also discuss the identification and estimation of this model. Importantly, in our empirical work, we will not consider the whole gamut of learning models, but restrict attention to models which are "close" to fully-rational in that the structure of the learning and decision rules are the same as in the benchmark fully-rational model (presented in Section 1.1 above); however, the rules themselves are allowed to be different.

We introduce the variable X_t^*, which denotes the subject's round t beliefs about the current state S_t; obviously, subjects know their beliefs X_t^*, but these are unobserved by the researcher.[16] 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:

[12] Studies by Camerer et al. (1993) and Johnson et al. (2002) (as well as the development of "Mouselab" system and its application to choice tasks in Payne et al., 1993) were the pioneering work in economics in which they applied mouse tracking in alternating-offer bargaining games. Then, the attention measures were employed to study forward induction (Camerer and Johnson, 2004), level-k models in matrix-games (Costa-Gomes et al., 2001), two-person guessing games (Costa-Gomes and Crawford, 2006), two-person games with private information about payoff-relevant states (Brocas et al., 2009), a boundedly rational, directed cognition model in goods choice tasks (Gabaix et al., 2006).

[13] Previous studies utilizing eye tracking to study strategic game-theoretic models also focused on the order in which subjects gazed at different areas to infer subjects' step-by-step, conscious thinking processes in more complicated game situations. Since our bandit choice problem is a relatively simple problem involving no strategic thinking, we do not focus on subjects' gaze orders in this paper.

[14] An alternative to using eye movements to proxy for beliefs would have been to elicit beliefs directly (as in Nyarko and Schotter, 2002). However, given the length of our experiments (8 blocks of 25 periods/trials each), and our need to have beliefs for each period, it seemed infeasible to elicit beliefs. Indeed, in some pilot experiments, we tried eliciting beliefs randomly after some periods, and found that this made the experiments unduly long. In addition, direct elicitation requires subjects make a conscious effort to input their subjective beliefs, while eye movements require less conscious and intentional effort, which keeps the bandit choice tasks more natural.

[15] The details on the discretization is in Appendix C.

[16] 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.

Assumption 1. Subjects' choice probabilities $P(Y_t|X_t^*)$ only depend on current beliefs. Moreover, the choice probabilities $P(Y_t = y|X_t^*)$ vary across different values of X_t^*: i.e.,

$$\Pr\left(Y_t = y \middle| X_t^* = \bar{x}\right) \neq \Pr\left(Y_t = y \middle| X_t^* = \tilde{x}\right) \quad \text{for } \bar{x} \neq \tilde{x}.$$

Because we interpret the unobserved variables X_t^* here as a reflection 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. This is justified by the simulation of the fully-rational decision rules under the reversal learning setting (in Section 1.1), which showed that these rules are myopic and depend only on current beliefs.[17]

This assumption embodies the core of our strategy for estimating subjects' beliefs; it posits important exclusion restrictions that, conditional on beliefs X_t^*, the observed action Y_t is independent of everything else, including the eye movement Z_t as well as past choices Y_{t-1}. Thus, beliefs (which are unobserved to the researcher) are the reason for serial correlation in choices observed in Table 2.[18]

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 fully-rational Bayesian belief-updating rule (cf. Eq. (4)), in which the period t beliefs depend only on the past beliefs, 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 $P(Z_t = i|Z_{t-1} = j)$, is invertible.
(iii) $E[Z_t|X_t^*]$ is increasing in X_t^*.

Assumption 3(i) is an important exclusion restriction that, conditional on X_t^*, the eye movement Z_t in period t is independent of Z_{t-1}. This assumption is reasonable because, in the experimental setup, 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.[19]

The invertibility assumption 3(ii) 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(iii) "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.

The crucial assertion underlying Assumption 3 is that our eye-movement measure Z is indeed a noisy measure of beliefs. Without this assertion, Assumptions 3(ii) and 3(iii) are violated, and the identification argument breaks down. While we have provided some evidence from both the existing literature as well as our experimental data to support this assertion in Section 1.3 above, we acknowledge that the link between gaze duration and beliefs may be somewhat weak.

Assumption 4. The conditional probability distributions describing subjects' choices $P(Y_t|X_t^*)$, learning rules $P(X_t^*|X_{t-1}^*, R_{t-1}, Y_{t-1})$, and eye movements $P(Z_t|X_t^*)$ are the same for all subjects and trials t.

This stationarity and homogeneity assumption[20] justifies pooling the data across all subjects and trials for estimating the model.

[17] Moreover, because we assume beliefs are discrete and take K values, there is a tension between the two parts of Assumption 1. The higher K is, the more reasonable is the first part, whereas the second part is more reasonable the lower K is. We thank a referee for pointing this out.

[18] Our nonparametric model also rules out the case that the precisions of subjects' beliefs affect their decisions. Accommodating this requires extending the nonparametric identification and estimation techniques utilized in this paper to incorporate a second latent variable, which is beyond the scope of this paper.

[19] 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^*. The results, which can be obtained from the authors upon request, indicate that the results are quite similar, for different values of Z_{t-1}, which imply that Assumption 3 is quite reasonable.

[20] In principle, this assumption could be relaxed with a much larger sample size; the subject homogeneity assumption can, in principle, be relaxed by gathering enough data per subject, such that the model could be estimated for each subject individually (see Wilcox, 2006, for a study of heterogeneity bias with experimental data). Given the eye fatigue facing subjects who are attached to an eye tracker, running so many trials per subject is not feasible.

2.1. Estimation and identification

In the model described previously, the unknown functions we want to estimate are:

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

Despite its simplicity, this model is not straightforward to estimate, because these unknown functions depend on the latent beliefs X_t^*, which are not only unobserved but changing over time.[21] The main results underlying estimation in this paper were established in Hu (2008), which considers the identification and estimation of nonlinear models with latent variables when auxiliary variables (which can be noisy proxies of the latent variables, such as the eye-tracking measures in the present paper) are available. This estimator is very simple, and involves only elementary calculations using matrices formed from the observed data.

Hu and Shum (2012) extend these results to dynamic Markovian models in which the latent variables can vary over time. The specific estimation procedure utilized in this paper is based on the simpler results in Hu (2008). It is possible to relax some of the assumptions from the previous section, and utilize the more involved results in Hu and Shum (2012) for estimation; however, we do not consider these possibilities here, because these results require additional periods of data, and our experimental sample is relatively limited.

For simplicity, we will use the shorthand notation $P(\cdots)$ to denote generically a probability distribution. The identification argument (and, subsequently, estimation procedure) takes two steps. In the first step, the goal is to recover the choice and eye-movement probability functions – that is, the probabilities $P(Y_t|X_t^*)$ (resp. $P(Z_t|X_t^*)$) of a given choice (resp. of given eye gaze duration) conditional on the latent beliefs. In the second step, we recover the learning rules. We describe both steps in turn.

First step. The joint probability distribution $P(Z_t, Y_t|Z_{t-1})$ can be factorized as follows:

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

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 $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are all $K \times K$ matrices, defined as:

$$\mathbf{A}_{y,Z_t|Z_{t-1}} = \big[P_{Y_t,Z_t|Z_{t-1}}(y,i|j)\big]_{i,j}; \qquad \mathbf{B}_{Z_t|X_t^*} = \big[P_{Z_t|X_t^*}(i|k)\big]_{i,k}; \qquad \mathbf{C}_{X_t^*|Z_{t-1}} = \big[P_{X_t^*|Z_{t-1}}(k|j)\big]_{k,j},$$

$$\mathbf{D}_{y|X_t^*} = \begin{bmatrix} P_{Y_t|X_t^*}(y|1) & 0 & 0 \\ 0 & P_{Y_t|X_t^*}(y|2) & 0 \\ & \ddots & \\ 0 & & 0 \\ 0 & 0 & P_{Y_t|X_t^*}(y|K) \end{bmatrix}. \tag{6}$$

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

$$\mathbf{G}_{Z_t|Z_{t-1}} = \big[P_{Z_t|Z_{t-1}}(i|j)\big]_{i,j}. \tag{7}$$

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

$$\mathbf{A}_{y,Z_t|Z_{t-1}} \mathbf{G}^{-1}_{Z_t|Z_{t-1}} = \mathbf{B}_{Z_t|X_t^*} \mathbf{D}_{y|X_t^*} \mathbf{B}^{-1}_{Z_t|X_t^*}. \tag{8}$$

Since $\mathbf{D}_{y|X_t^*}$ is a diagonal matrix, this equation represents an eigenvalue decomposition of the matrix $\mathbf{A}_{y,Z_t|Z_{t-1}} \mathbf{G}^{-1}_{Z_t|Z_{t-1}}$, which can be computed from the observed data sequence $\{Y_t, Z_t\}$.[22] This shows that from the observed data, we can identify the

[21] Specifically, this model is a nonlinear "hidden state Markov" model, which is typically quite challenging to estimate (cf. Ghahramani, 2001).
[22] From Eq. (7), the invertibility of $\mathbf{G}$ (which is Assumption 3(i)) implies the invertibility of $\mathbf{B}$.

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

In order for this argument to be valid, the eigendecomposition in Eq. (8) must be unique. This requires the eigenvalues (corresponding to choice probabilities $P(y|X_t^*)$) to be distinctive; that is, $P(y|X_t^*)$ should vary in X_t^* – which Assumption 1 ensures. Furthermore, the eigendecomposition in Eq. (8) 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 correspond to the conditional probabilities $P(Z_t|X_t^*)$, it is appropriate to normalize each column so that it sums to one.

Second step. We begin by factorizing the conditional probability distribution

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

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

$$\mathbf{H}_{Z_{t+1},y,r,Z_t} = \big[P_{Z_{t+1},Y_t,R_t,Z_t}(i,y,r,j)\big]_{i,j}, \qquad \mathbf{L}_{X_{t+1}^*,X_t^*,y,r} = \big[P_{X_{t+1}^*,X_t^*,Y_t,R_t}(i,j,y,r)\big]_{i,j}. \tag{9}$$

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

$$\mathbf{L}_{X_{t+1}^*,X_t^*,y,r} = \mathbf{B}^{-1}_{Z_{t+1}|X_{t+1}^*}\mathbf{H}_{Z_{t+1},y,r,Z_t}\big[\mathbf{B}'_{Z_t|X_t^*}\big]^{-1}. \tag{10}$$

This result implies that two consecutive periods of experimental and eye-movement data $(Z_t, Y_t, R_t), (Z_{t-1}, Y_{t-1}, R_{t-1})$ from each subject suffice to identify and estimate the decision and learning rules in this model.

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. (8) and (10). To implement this, we assume that the eye-movement measures Z_t and the unobserved beliefs X_t^* are discrete, and take three values.[23]

Moreover, 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 $\tau = 2, \ldots, 25$.[24] Results from Monte Carlo simulations (available from the authors on request) show that the estimation procedure produces accurate estimates of the model components.[25]

3. Results

Tables 3 and 4 contain results. The beliefs X_t^* are assumed to take the three 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 "indifferent".[26] Accordingly, the eye movements Z_t were also discretized to three values, as discussed before.

Table 3 contains the estimates of the choice and eye-movement probabilities. The first and last columns of the panels in this table indicate that choices and eye movements are closely aligned with beliefs, when beliefs are sufficiently strong (i.e. are equal to either $X^* = 1$ or $X^* = 3$). Specifically, in these results, the probability of choosing a color contrary to beliefs –

[23] The details concerning the discretization of the eye-movement measure Z_t are given in Appendix C.

[24] 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; the ergodic theorem then 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.

[25] Moreover, 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. (8) and (10), 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.

[26] We also estimated the model allowing for more belief states ($\geqslant 4$), but the results we obtained, while qualitatively similar to the previous results, were quite imprecise. This arises from our nonparametric approach, for which it is difficult to obtain reliable estimates with modest sample sizes.

Table 3
Estimates of choice and eye-movement probabilities.

Estimated choice probabilities: $P(Y_t \mid X_t^*)$			
X_t^*:	1 (green)	2 (indifferent)	3 (blue)
$Y_t = 1$ (green)	0.9866	0.4421	0.0064
	(0.0254)	(0.1012)	(0.0145)
2 (blue)	0.0134	0.5579	0.9936

Estimated eye-movement probabilities: $P(Z_t \mid X_t^*)$			
X_t^*:	1 (green)	2 (indifferent)	3 (blue)
$Z_t = 1$ (green)	0.8639	0.2189	0.0599
	(0.0444)	(0.1301)	(0.0529)
2 (middle)	0.0815	0.6311	0.0980
	(0.0526)	(0.1745)	(0.0455)
3 (blue)	0.0546	0.1499	0.8421
	(0.0400)	(0.0949)	(0.0939)

Each cell contains parameter estimates, with block-bootstrapped (subjects drawn with replacement) standard errors in parentheses. Each column sums to one.

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 $X_t^* = 3$.[27]

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 indifferent, they tend to split their gaze more evenly between the two colors (i.e. $Z_t = 2$) 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 (i.e. choices made in accordance with beliefs) are taken, and illustrate an important asymmetry. 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 beliefs 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 ($R_1 = 1$). This muted updating following failures[28] 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" (contrary to current beliefs) choices. 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%). 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 Fig. 3 show that explorative choices occur with very low probability, leading to imprecision in the estimates of belief-updating rules following such choices.

The middle 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 ($X_t^* = 2$), subjects update more towards "green" when a previous choice of green yielded the high rather than the low 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" ($Y_t = 2$).

3.1. Comparing choice and learning rules across different models

Next, we compare our estimated learning rules to alternative learning rules which have been considered in the literature. We consider three alternative learning rules: (i) the *fully-rational Bayesian* model, which is the model discussed in Section 1.1 above; (ii) *reinforcement learning* (cf. Sutton and Barto, 1998); and (iii) *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 are given in Appendix B.

[27] We also considered a robustness check against the possibility that subjects' gazes 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.

[28] The reluctance to give up bad choices seems to suggest some form of "anticipated regret"; however, regret matters when outcomes of unchosen options are revealed to decision-makers, leading to regret at not choosing one of these unchosen options. This is not a feature of our setting. The reluctance to give up bad choices is also related to representativeness (Kahneman and Tversky, 1972) and conservatism (Edwards, 1968) in experimental psychology, where people underreact to signals which are not representative of their beliefs. We thank a referee for bringing these theories to our attention.

Table 4
Estimates of learning (belief-updating) rules.

$P(X^*_{t+1}\|X^*_t, y, r)$, $r=1$ (lose), $y=1$ (green)			
X^*_t:	1 (green)	2 (indifferent)	3 (blue)
$X^*_{t+1}=1$ (green)	0.5724	0.3075	0.1779
	(0.1032)	(0.0871)	(0.2161)
2 (indifferent)	0.0000[a]	0.3138	0.4002
	(0.0818)	(0.0977)	(0.1885)
3 (blue)	0.4276	0.3787	0.4219
	(0.0920)	(0.0841)	(0.2264)

$P(X^*_{t+1}\|X^*_t, y, r)$, $r=2$ (win), $y=1$ (green)			
X^*_t:	1 (green)	2 (indifferent)	3 (blue)
$X^*_{t+1}=1$ (green)	0.8889	0.6621	0.8242
	(0.0890)	(0.1234)	(0.2372)
2 (indifferent)	0.0000	0.2702	0.1758
	(0.0667)	(0.1131)	(0.1821)
3 (blue)	0.1111	0.0678	0.0000
	(0.0654)	(0.0510)	(0.1418)

$P(X^*_{t+1}\|X^*_t, y, r)$, $r=1$ (lose), $y=2$ (blue)			
X^*_t:	3 (blue)	2 (indifferent)	1 (green)
$X^*_{t+1}=3$ (blue)	0.5376	0.2297	0.2123
	(0.0848)	(0.0929)	(0.1621)
2 (indifferent)	0.0458	0.2096	0.1086
	(0.0803)	(0.1167)	(0.1814)
1 (green)	0.4166	0.5607	0.6792
	(0.0764)	(0.1291)	(0.2231)

$P(X^*_{t+1}\|X^*_t, y, r)$, $r=2$ (win), $y=2$ (blue)			
X^*_t:	3 (blue)	2 (indifferent)	1 (green)
$X^*_{t+1}=3$ (blue)	0.8845	0.6163	0.6319
	(0.0918)	(0.0993)	(0.1794)
2 (indifferent)	0.0000	0.3558	0.3566
	(0.0909)	(0.1106)	(0.1753)
1 (green)	0.1155	0.0279	0.0116
	(0.0454)	(0.0466)	(0.1046)

Each cell contains parameter estimates, with block-bootstrapped (subjects drawn with replacement) standard errors in parentheses. Each column sums to one.

[a] 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 footnote 25.

Table 5
Simulated payoffs from learning models.

	Fully-rational Bayesian	Nonparametric	Reinforcement learning	Win-stay
5-%tile	$5	$1	$1	$1
25-%tile	$12	$8	$8	$8
50-%tile	$17	$13	$13	$13
75-%tile	$22	$18	$18	$18
95-%tile	$29	$25	$25	$25

The fully-rational model is described in Section 1.1, while the reinforcement learning 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) are reported.

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 subjects. The empirical quantiles of these distributions are presented in Table 5.

As we expect, the fully-rational Bayesian model generates the most revenue for subjects; the 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 (net of the fixed show-up fee).

Next, we seek 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. For the fully-rational Bayesian and

Table 6
Summary statistics for beliefs in three learning models. X^*: Beliefs from nonparametric model. B^*: Beliefs from fully-rational Bayesian model. V^*: "Beliefs" (valuations) from reinforcement learning model.

Panel 1: Belief frequency in nonparametric model

X^*	1 (green)	2 (indifferent)	3 (blue)
	1878 (45%)	366 (10%)	1956 (45%)

Panel 2: Beliefs from other models

	Mean	Median	Std.	Lower 33%	Upper 33%
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

All three measures of beliefs are oriented so that larger values correspond to a more favorable assessment that "blue" is currently the good arm. See Appendix B for details on computation of beliefs in these three learning models.

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.[29]

Table 6 contains 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.

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 "indifferent" 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 summarizes the beliefs from the reinforcement learning and Bayesian models. The reinforcement learning valuation measure V^* appears symmetric and centered around zero, while the average Bayesian belief 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 different 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.[30] These are shown in Table 7.

The most striking difference between the three sets of learning rules lies in how beliefs update following unsuccessful choices (i.e. choices which yielded a negative reward). Comparing the Bayesian and the nonparametric learning rules (in Table 4), we see that Bayesian beliefs exhibit less "stickiness" 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 rule estimates (Table 4) 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: the nonparametric learning rule continues to place 54% probability on blue, whereas the fully-rational Bayesian belief is only 30%.[31]

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, then, 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 fully-rational model may explain the lower revenue generated by these models.

In Table 8 we compare the choice rules across the different models. As before, we discretized the beliefs from each model into three values. Overall, the choice rules from the nonparametric model, in Table 3, are quite close to the fully-rational model. This suggests that the lower payoffs from the nonparametric model relative to the fully-rational model arise primarily

[29] There are no beliefs in the win-stay model, which is a simple choice heuristic. The Bayesian beliefs B_t^* are obtained from Eq. (4) and evaluated at the observed sequence of choices and rewards (Y_t, R_t). Appendix B contains additional details on how the beliefs were derived for the other models.

[30] 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.

[31] This finding that subjects are reluctant to update down after unsuccessful choices appears somewhat novel. It goes against previous experimental findings in Meyer and Shi (1995) (who find that subjects tend to switch too often in two-armed bandit problems), and also contrasts field findings in Strahilevitz et al. (2011), that investors tend to shun stocks which they had previously sold for a loss.

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

$P(X^*_{t+1}\|X^*_t, y, r)$, $r=1$ (lose), $y=1$ (green)						
Beliefs B^*_{t+1}, V^*_{t+1}:	Fully-rational Bayesian			Reinforcement learning		
	1 (green)	2 (indifferent)	3 (blue)	1 (green)	2 (indifferent)	3 (blue)
1 (green)	0.2878	0	0	0.6538	0	0
2 (indifferent)	0.1730	0	0	0.1381	0.0115	0
3 (blue)	0.5392	1.0000	1.0000	0.2080	0.9885	1.0000
$P(X^*_{t+1}\|X^*_t, y, r)$, $r=2$ (win), $y=1$ (green)						
Beliefs B^*_{t+1}, V^*_{t+1}:	Fully-rational Bayesian			Reinforcement learning		
	1 (green)	2 (indifferent)	3 (blue)	1 (green)	2 (indifferent)	3 (blue)
1 (green)	1.0000	1.0000	0.6734	1.0000	0.8818	0.6652
2 (indifferent)	0	0	0.1250	0	0.1182	0.1674
3 (blue)	0	0	0.2016	0	0	0.1674
$P(X^*_{t+1}\|X^*_t, y, r)$, $r=1$ (lose), $y=2$ (blue)						
Beliefs B^*_{t+1}, V^*_{t+1}:	Fully-rational Bayesian			Reinforcement learning		
	3 (blue)	2 (indifferent)	1 (green)	3 (blue)	2 (indifferent)	1 (green)
3 (blue)	0.3060	0	0	0.6576	0	0
2 (indifferent)	0.1601	0	0	0.1261	0.0109	0
1 (green)	0.5338	1.0000	1.0000	0.2164	0.9891	1.0000
$P(X^*_{t+1}\|X^*_t, y, r)$, $r=2$ (win), $y=2$ (blue)						
Beliefs B^*_{t+1}, V^*_{t+1}:	Fully-rational Bayesian			Reinforcement learning		
	3 (blue)	2 (indifferent)	1 (green)	3 (blue)	2 (indifferent)	1 (green)
3 (blue)	1.0000	1.0000	0.6760	1.0000	0.8898	0.6983
2 (indifferent)	0	0.0000	0.1440	0	0.1102	0.1379
1 (green)	0	0	0.1800	0	0	0.1638

All three measures of beliefs are oriented so that larger values correspond to a more favorable assessment that "blue" is currently the good arm.

Table 8
Choice probabilities for alternative learning models.

Fully-rational Bayesian			
Beliefs B^*_t:	1 (green)	2 (indifferent)	3 (blue)
$Y_t=1$ (green)	1.0000	0.5000	0.0000
2 (blue)	0.0000	0.5000	1.0000
Reinforcement learning			
Beliefs V^*_t:	1 (green)	2 (indifferent)	3 (blue)
$Y_t=1$ (green)	0.7629	0.4939	0.2250
2 (blue)	0.2371	0.5061	0.7750

All three measures of beliefs are oriented so that larger values correspond to a more favorable assessment that "blue" is currently the good arm.

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. The choice rules are much smoother than in the fully-rational Bayesian model and the estimated model. This suggests that the similarities of the payoffs from the nonparametric and reinforcement learning models (as shown in Table 5) are due to the similarities in belief-updating rules, and not to the choice rules, which are quite different in the two models. In addition, 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.

4. Conclusions

We estimate learning rules from data drawn from experiments of multi-armed bandit problems. The experimental data are augmented by measurements of subjects' eye movements, which proxy for their beliefs. Subjects' payoffs from following the estimated learning rules are lower than what would be obtained from the fully-rational Bayesian model, and comparable to the profits obtained from a reinforcement learning model, and a win-stay choice heuristic. Relatively to the fully-rational

model, the belief-updating rules estimated for the nonparametric models show that subjects appear reluctant to "update down" following unsuccessful choices, leading to the sub-optimality of this rule (in terms of profits).

Our nonparametric estimator for subjects' choice probabilities and learning rules is easy to implement, involving only elementary matrix operations. Furthermore, the nonparametric estimation of learning models using experimental data appears to be a portable idea which may be applied broadly to other experiments involving dynamic decision problems.

There are several caveats of our analysis which point the way to future work. First, because our estimation procedure is nonparametric, large sample sizes are needed to produce reliable estimates, which necessitated to pooling of data across subjects. Future research will focus on ways to allow for more heterogeneity across subjects. Second, our identification procedure relies crucially on the existence of a noisy auxiliary measure of beliefs – in this case the gaze duration measure Z. The existing evidence on the link between gaze duration and beliefs is spotty, and in future work we will seek out alternative auxiliary measures of beliefs which may be more robust.

Appendix A. Details for computing fully-rational Bayesian learning model

Here we provide more details about the simulation of the fully-rational model from Section 1.1. The Bellman equation corresponding to the dynamic optimization problem is:

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

The state variable in this model is B_t^*, the beliefs at the beginning of each period. Above, the expectation $E_{B_{t+1}^*|B_t^*,Y_t,R_t}$ is taken with respect to Eq. (4), 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

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

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:

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

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 $\hat{y}_T^*(B) = \arg\max_i \mathbb{E}[R_t|i,B]$ for each value of $B\in\mathcal{B}$. Subsequently, proceeding backwards, we can obtain the value and choice functions for periods $t = T-1, T-2,\ldots,1$. The resulting choice functions are described in Section 2.1.

Appendix B. Details on model fitting in alternative learning models

In Section 3.1, we compared belief dynamics in the nonparametric model (X^*) with counterparts in other dynamic choice models. Here we provide additional details on how these quantities were computing for each model.

Recovering 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 eye-movement 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

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

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:

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

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 "indifferent" at the beginning of each block). To facilitate this optimization problem, we derive the best-fitting 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.[32]

Reinforcement learning model. We employ a TD (Temporal-Difference)-Learning models (Sutton and Barto, 1998, Section 6) in which action values are updated via the Rescorla–Wagner rule. The value updating rule for this model is given by:

$$V_{Y_t}^{t+1} \longleftarrow V_{Y_t}^{t} + \alpha\delta_t, \tag{14}$$

where Y_t denotes the choice taken in trial t, α denotes the learning rate, and δ_t denotes the "prediction error" δ_t for trial t, defined as:

$$\delta_t = R_t - V_{Y_t}^{t}, \tag{15}$$

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 "softmax" (i.e. logit) form with the smoothing (or "temperature") parameter τ:

$$P_c^t = e^{V_c^t/\tau} \Big/ \left[\sum_{c'} e^{V_{c'}^t/\tau}\right]. \tag{16}$$

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

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

We plug in these values into Eqs. (14), (15), and (16) 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

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

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.

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 δ). 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 $1 - \epsilon$. The estimates we obtained for δ and ϵ, using maximum likelihood are: $\delta = 0.1268$ (0.0142); $\epsilon = 0.4994$ (0.0213).

Appendix C. Details on discretization of eye movements

Fig. 2 contains a histogram of the undiscretized eye-movement measure $\tilde{Z}_t$. It is clearly trimodal, with peaks at -1, 0 and 1, suggesting that a three-value discretization sufficiently captures most of that variation in eye movements. In the empirical work, we use the following three-value discretization:

$$Z_t = \begin{cases} 1 & \text{if } \tilde{Z}_t < -\sigma_z, \\ 2 & \text{if } -\sigma_z \leqslant \tilde{Z}_t \leqslant \sigma_z, \\ 3 & \text{if } \sigma_z < \tilde{Z}_t \end{cases} \tag{19}$$

where σ_z, a discretizing constant, is set to $\sigma_z = 0.20$.[33]

Appendix D. Hit rate, left choice rate and blue choice rate

The left panel in Fig. 3 shows the hit rates for each subject. Most of them are between 0.50 and 0.60, suggesting that there seems no radical differences. The middle panel shows the frequency of the machine choices on left side of the screen.

[32] 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 $\mathrm{Corr}(Y_t, X_t^*) = 0.99$, and in practically all periods, the estimated beliefs and observed choices coincided (i.e. $X_t^* = Y_t$). However, we felt that this did not accurately reflect subjects' beliefs.

[33] We find little difference in the estimation results if we vary σ_z from 0.05 to around 0.40, suggesting that the model is robust for different values of σ_z.

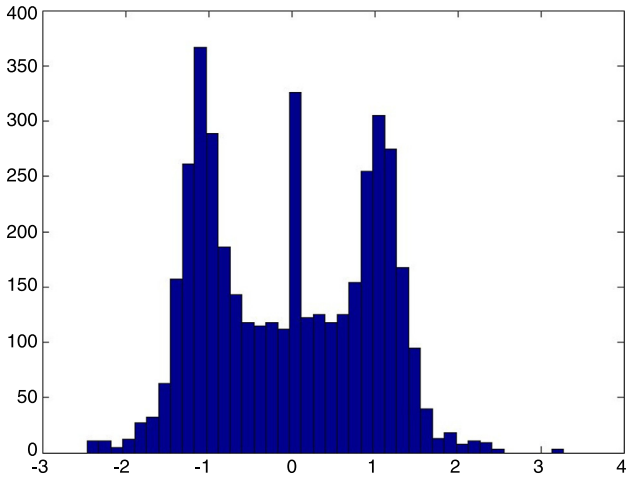

Fig. 2. Histogram of undiscretized eye-movement measure $\tilde{Z}_t$.

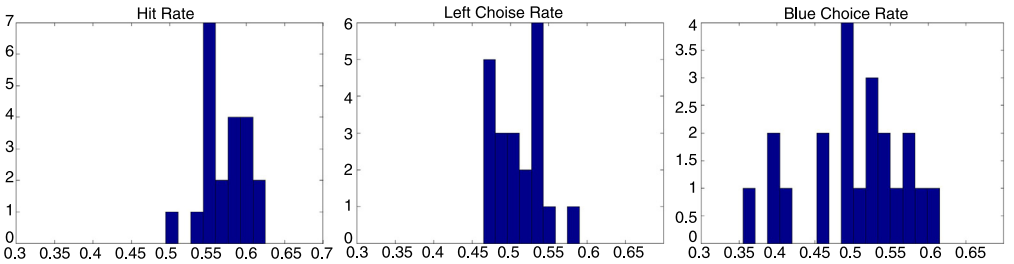

Fig. 3. Histogram: hit rate, blue choice rate and left choice rate.

The rates are around 0.5, suggesting that each subject has no specific preference for a certain color. The right panel shows the frequency of blue machine choices. Overall, the values are around 0.5, suggesting that these subjects have no systemic preference for a certain color.

References

Ackerberg, D., 2003. Advertising, learning, and consumer choice in experience good markets: A structural examination. Int. Econ. Rev. 44, 1007–1040.

Anderson, C., 2012. Ambiguity aversion in multi-armed bandit problems. Theory Dec. 72 (1), 15–33.

Armel, K., Rangel, A., 2008. The impact of computation time and experience on decision values. Amer. Econ. Rev. 98 (2), 163–168.

Armel, K., Beaumel, A., Rangel, A., 2008. Biasing simple choices by manipulating relative visual attention. Judgm. Dec. Making 3 (5), 396–403.

Bajari, P., Hortaçsu, A., 2005. Are structural estimates of auction models reasonable? Evidence from experimental data. J. Polit. Economy 113, 703–741.

Banks, J., Olson, M., Porter, D., 1997. An experimental analysis of the bandit problem. Econ. Theory 10 (1), 55–77.

Bergemann, D., Hege, U., 1998. Venture capital financing, moral hazard, and learning. J. Banking Finance 22 (6), 703–735.

Boorman, E., Behrens, T., Woolrich, M., Rushworth, M., 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., Carrillo, J., Wang, S., Camerer, C., 2009. Measuring attention and strategic behavior in games with private information. Mimeo, USC.

Brown, A.L., Camerer, C.F., Lovallo, D., 2012. To review or not to review? Limited strategic thinking at the box office. Amer. Econ. J. Microecon 4 (2), 1–28.

Camerer, C., Johnson, E., 2004. Thinking about attention in games: Backward and forward induction. In: Brocas, I., Carrillo, J. (Eds.), The Psychology of Economic Decisions, vol. 2: Reasons and Choices.

Camerer, C., Johnson, E., Rymon, T., Sen, S., 1993. Cognition and framing in sequential bargaining for gains and losses. In: Frontiers of Game Theory, pp. 27–47.

Caplin, A., Dean, M., 2008. Economic insights from 'neuroeconomic' data. Amer. Econ. Rev. 98 (2), 169–174.

Caplin, A., Dean, M., Glimcher, P., Rutledge, R., 2010. Measuring beliefs and rewards: A neuroeconomic approach. Quart. J. Econ. 125, 923–960.

Chan, T., Hamilton, B., 2006. Learning, private information, and the economic evaluation of randomized experiments. J. Polit. Economy 114, 997–1040.

Charness, G., Levin, D., 2005. When optimal choices feel wrong: A laboratory study of Bayesian updating, Complexity, and Affect. Amer. Econ. Rev. 95, 1300–1309.

Costa-Gomes, M., Crawford, V., 2006. Cognition and behavior in two-person guessing games: An experimental study. Amer. Econ. Rev. 96 (5), 1737–1768.

Costa-Gomes, M., Crawford, V., Broseta, B., 2001. Cognition and behavior in normal-form games: An experimental study. Econometrica 69 (5), 1193–1235.

Crawford, G., Shum, M., 2005. Uncertainty and learning in pharmaceutical demand. Econometrica 73, 1137–1174.

Crawford, V., Iriberri, N., 2007. Level-*k* auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? Econometrica 75 (6), 1721–1770.

Edwards, W., 1968. Conservatism in human information processing. In: Formal Representation of Human Judgment, pp. 17–52.

El-Gamal, M., Grether, D., 1995. Are people Bayesian? Uncovering behavioral strategies. J. Amer. Statistical Assoc. 90, 1137–1145.

Erdem, T., Keane, M., 1996. Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. Marketing Sci. 15, 1–20.

Gabaix, X., Laibson, D., Moloche, G., Weinberg, S., 2006. Costly information acquisition: Experimental analysis of a boundedly rational model. Amer. Econ. Rev. 96 (4), 1043–1068.
Gans, N., Knox, G., Croson, R., 2007. Simple models of discrete choice and their performance in bandit experiments. Manufact. Service Operations Manage. 9 (4), 383–408.
Ghahramani, Z., 2001. An introduction to hidden Markov models and Bayesian networks. Int. J. Pattern Recogn. Artificial Intel. 15, 9–42.
Gillen, B., 2011. Identification and estimation of level-*k* auctions. Mimeo, Caltech.
Gittins, J., Jones, G., 1974. A dynamic allocation index for the sequential design of experiments. In: Gani, J., et al. (Eds.), Progress in Statistics. North-Holland.
Goldfarb, A., Xiao, M., 2011. Who thinks about the competition? Managerial ability and strategic entry in US local telephone markets. Amer. Econ. Rev. 101, 3130–3161.
Hampton, A., Bossaerts, P., O'Doherty, J., 2006. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J. Neurosci. 26, 8360–8367.
Hikosaka, O., Nakamura, K., Nakahara, H., 2006. Basal ganglia orient eyes to reward. J. Neurophysiol. 95 (2), 567.
Ho, T., Su, X., 2010. A dynamic level-*k* model in games. Mimeo, University of California at Berkeley.
Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., Camerer, C., 2005. Neural systems responding to degrees of uncertainty in human decision-making. Science 310 (5754), 1680–1683.
Hsu, M., Krajbich, I., Zhao, C., Camerer, C., 2009. Neural response to reward anticipation under risk is nonlinear in probabilities. J. Neurosci. 29 (7), 2231–2237.
Hu, Y., 2008. Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. J. Econometrics 144, 27–61.
Hu, Y., Shum, M., 2012. Nonparametric identification of dynamic models with unobserved state variables. J. Econometrics 171, 32–44.
Johnson, E., Camerer, C., Sen, S., Rymon, T., 2002. Detecting failures of backward induction: Monitoring information search in sequential bargaining. J. Econ. Theory 104 (1), 16–47.
Jovanovic, B., 1979. Job matching and the theory of turnover. J. Polit. Economy, 972–990.
Kahneman, D., Tversky, A., 1972. Subjective probability: A judgment of representativeness. Cogn. Psychol. 3 (3), 430–454.
Kawagoe, R., Takikawa, Y., Hikosaka, O., 1998. Expectation of reward modulates cognitive signals in the basal ganglia. Nature Neurosci. 1, 411–416.
Knoepfle, D., Wang, J., Camerer, C., 2009. Studying learning in games using eye-tracking. J. Europ. Econ. Assoc. 7 (2–3), 388–398.
Krajbich, I., Armel, C., Rangel, A., 2010. Visual fixations and the computation and comparison of value in simple choice. Nature Neurosci. 13, 1292–1298.
Kuhnen, C., Knutson, B., 2008. The influence of affect on beliefs, preferences and financial decisions. MPRA Paper 10410. University Library of Munich, Germany.
Lauwereyns, J., Watanabe, K., Coe, B., Hikosaka, O., 2002. A neural correlate of response bias in monkey caudate nucleus. Nature 418, 413–417.
Meyer, R., Shi, Y., 1995. Sequential choice under ambiguity: Intuitive solutions to the armed-bandit problem. Manage. Sci. 41 (5), 817–834.
Miller, R., 1984. Job matching and occupational choice. J. Polit. Economy 92, 1086–1120.
Nyarko, Y., Schotter, A., 2002. An experimental study of belief learning using elicited beliefs. Econometrica 70, 971–1005.
Payne, J., Bettman, J., Johnson, E., 1993. The Adaptive Decision Maker. Cambridge University Press.
Payzan-LeNestour, E., Bossaerts, P., 2011. Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. PLoS Computational Biol. 7 (1), 1704–1711.
Preuschoff, K., Bossaerts, P., Quartz, S., 2006. Neural differentiation of expected reward and risk in human subcortical structures. Neuron 51 (3), 381–390.
Preuschoff, K., Hart, B. Marius, Einhauser, W., 2011. Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. Front. Dec. Neurosci. 5.
Reutskaja, E., Nagel, R., Camerer, C., Rangel, A., 2011. Search dynamics in consumer choice under time pressure: An eye-tracking study. Amer. Econ. Rev. 101 (2), 900–926.
Rothschild, M., 1974. A two-armed bandit theory of market pricing. J. Econ. Theory 9 (2), 185–202.
Shimojo, S., Simion, C., Shimojo, E., Scheier, C., 2003. Gaze bias both reflects and influences preference. Nature Neurosci. 6 (12), 1317–1322.
Sokol-Hessner, P., Hsu, M., Curley, N., Delgado, M., Camerer, C., Phelps, E., 2009. Thinking like a trader selectively reduces individuals' loss aversion. Proc. Natl. Acad. Sci. 106 (13), 5035.
Strahilevitz, M., Odean, T., Barber, B., 2011. Once burned, twice shy: How naïve learning, counterfactuals, and regret affect the repurchase of stocks previously sold. J. Marketing Res. 48, 102–120.
Sutton, R., Barto, A., 1998. Reinforcement Learning. MIT Press.
Wang, J., Spezio, M., Camerer, C., 2010. Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in sender–receiver games. Amer. Econ. Rev. 100 (3), 984–1007.
Weitzman, M., 1979. Optimal search for the best alternative. Econometrica, 641–654.
Wilcox, N., 2006. Theories of learning in games and heterogeneity bias. Econometrica 74, 1271–1292.