Multi-state choices with aggregate feedback on unfamiliar alternatives

Philippe Jehiel * and Juni Singh †

February 26, 2021

Abstract

This paper studies a multi-state binary choice experiment in which in each state, one alternative has well understood consequences whereas the other alternative has unknown consequences. Subjects repeatedly receive feedback from past choices about the consequences of unfamiliar alternatives but this feedback is aggregated over states. Varying the payoffs attached to the various alternatives in various states allows us to test whether unfamiliar alternatives are discounted and whether subjects' use of feedback is better explained by similarity-based reinforcement learning models (in the spirit of the valuation equilibrium, Jehiel and Samet 2007) or by some variant of Bayesian learning model. Our experimental data suggest that there is no discount attached to the unfamiliar alternatives and that similarity-based reinforcement learning models have a better explanatory power than their Bayesian counterparts.

Key words: Ambiguity, Bounded Rationality, Experiment, Learning, Coarse feedback, Valuation equilibrium

JEL Classification: D81, D83, C12, C91

We are grateful to PSE Research grants, LABEX OSE as well as the ERC grant no. 742816 for funding. Special thanks to Maxim Frolov for helping us in the execution of the experiment and to Philipp Ketz for helping us with the econometric analysis. We have also benefited from the comments of Arian Charpin, Elias Bouacida, Emmanuel Vespa, Guillaume Frechette, Itzhak Gilboa, Jean-Francois Laslier, Jean-Marc Tallon, Julien Combe, Nicolas Gefflot, Nicolas Jacquemet and Peyton Young. We also thank the editorial team for useful comments.

*Paris School of Economics and University College of London; email: jehiel@enpc.fr

†Paris School of Economics

1 Introduction

In many situations, the decision maker faces a choice between two alternatives one of them being more familiar and thus easier to evaluate and another one being less familiar and thus harder to assess. There is generally some information about the less familiar alternative, but this information is typically coarse not being entirely relevant to the specific context of interest.

To give a concrete application, think of the adoption of a new technology by farmers. A farmer has a lot of information about the performance of the current technology but not so much about the new one. The farmer may collect information about the new technology by asking around other farmers who would have previously adopted it. But due to the heterogeneity of the soil and/or the heterogeneity in the ability of the farmers, what works well/poorly for one farmer need not perform in the same way for another. Thus, the feedback received about the new technology is coarse in the sense that it is aggregated over different situations (states in the decision theoretic terminology) as compared to the information held for the old technology.[1] Another example may concern hiring decisions.[2] Consider hiring for two different jobs, one requiring high skill going together with higher education level and the other requiring lower skills, and assume potential candidates either come from a majority group or a minority group (as determined by nationality, color, caste or religion, say). Presumably, there is a lot of familiarity with the majority group allowing for this group to distinguish the productivity as a function of education as well as past experiences. However, for the minority group, information is more likely to be coarse and perceived productivity in that group may not be as easy to relate to education or past experiences.

We are interested in understanding how decision makers would make their decisions in multi-state binary decision problems in which they have precise state-specific information about the performance of one alternative and less precise information about the other. The less precise information takes the form that the decision maker receives aggregate (not state-specific) feedback about the performance of that alternative. Our interest lies in understanding the behavior of agents acting repeatedly in such environments.

To shed light on this, we have conducted experiments in which a pool of subjects have to choose between two actions over several rounds. In each round and for each subject, the choice is to be made in one of two states with an equal proportion for the two states across rounds and across subjects. In state 1, a subject has to choose between a blue action and a red action. In state 2, a subject has to choose between a green action and a

[1] Ryan and Gross (1946) propose an early study of the diffusion of new technology adoption in the farming context. See also Young (2009) for a study focused on the diffusion dimension.

[2] This example is inspired by Fryer and Jackson (2008)'s discussion of discrimination and categorization.

red action. All actions give stochastic returns. While the return distribution to the blue and green actions are known with some precision from the start, the returns to the red actions are not, and importantly the returns to these actions may be different in the two states. As subjects make choices, they receive feedback about the returns to the various actions (chosen in the last round in the entire pool) as a function of their color. Thus, the feedback for the return to a red action is coarse to the extent that it is not known whether the red action was chosen in state 1 or in state 2. We consider different treatments in which we vary the distribution of the returns to the two red actions while keeping the same return characteristics for the blue and green actions so that the initial conditions for each of these treatments are identical. In each treatment, we are interested in the evolution of the choice patterns across rounds and whether these stabilize after a sufficiently large number of rounds. That is, we aim at shedding light on the type of learning model that seems most appropriate to describe the evolution of behaviors in such contexts with coarse feedback, and also on the type of equilibrium concept to be used in such contexts (that could be relevant to describe patterns of behaviors after they stabilize).

Before presenting our main results, let us mention several possible theoretical benchmarks that could be relevant for our study. First, in the tradition of reinforcement learning (see Barto and Sutton 1998 or Fudenberg and Levine 1998 for textbook expositions), subjects could assess the strength of actions by considering their average return observed in the past. One key difficulty in our context is that for the red actions, the feedback is coarse not disentangling the corresponding return whether the red action was chosen in state 1 or 2. Following Jehiel and Samet (2007), one could extend such an approach by considering a similarity-based reinforcement learning model in which a single valuation would be attached to the two red actions (and reinforced accordingly) and the two red actions would be considered alike in terms of strength by the learning subjects. Jehiel and Samet (2007) have proposed a solution concept called the valuation equilibrium aimed at capturing the limiting outcomes of such similarity-based reinforcement learning models. We describe an extension of this solution concept allowing subjects to rely on noisy best-responses in the vein of the logit model (as popularized by McKelvey and Palfrey (1995) in experimental economics). A key feature of the valuation equilibrium is that the valuation of the red actions is endogenously shaped by the relative frequency with which the red actions are chosen in the two states: When the red action is much more frequently chosen in state 2 (resp. 1) than in state 1 (resp. 2), the induced valuation for red is much closer to the return to the red action in state 2 (resp. 1). When the red actions are chosen with the same frequency in the two states, the valuation of red is just the unweighted average of the returns to the two red actions. The returns to the red actions in the three treatments were chosen to give rise to different averaging in the determination of the valuation of the Red actions as well different predictions for the valuation equilibrium and the optimal solution.

Second, subjects could form beliefs about the returns to the red actions in the two states relying on some form of Bayesian updating to adjust the beliefs after they get addi-

tional feedback. Note that in our experiment, subjects get to observe the number of times the blue and green actions were chosen in the last round.This together with the knowledge that the two states are equally likely is informative whether the coarse feedback observed for the red actions is more representative of state 1 or 2 (for example a strong imbalance in favor of the green actions as opposed to the blue actions would be indicative that the previous red choices corresponded more to state 1 where the blue action was available). Of course, such a Bayesian approach heavily depends on the initial prior, and when studying such models, we will consider several families of priors. For the reinforcement learning model we will assume that subjects employ a noisy best response of the logit type.

Another key theoretical consideration is that the feedback concerning the red actions is ambiguous to the extent that it does not distinguish between the returns to the red actions according to the state $s = 1$ or 2. Following the tradition of Ellsberg (1961), one may suspect then that subjects would apply an ambiguity discount to the red actions (see Gilboa and Schmeidler (1989) for an axiomatization of ambiguity aversion). In the terminology of Epstein and Schneider (2007) or Epstein and Halevy (2019), the coarse feedback about the red actions can be viewed as an ambiguous signal. To cope with the ambiguous nature of the feedback in a simple way, we propose adding to the previous models (the similarity-based reinforcement learning and the Bayesian model) an ambiguity discount to the assessment of the red actions.

Based on our observed experimental data, we estimate both for the similarity-based reinforcement learning model and the Bayesian learning model the parameters that fit our experimental data best. A question of is interest whether there is a discounting applied to the red actions and for which learning model. After performing these estimations, we ask our main question of interest: which of the similarity-based reinforcement learning or the generalized Bayesian learning model explains the observed data best. We also study if the observed choice patterns in the final rounds of our experiments well explained by the valuation equilibrium.

Our main findings are as follows. First, in our estimation of the similarity-based reinforcement learning model, we find that there is no ambiguity discount. That is, despite the inherent ambiguity of the feedback received about the red actions, the red actions are not discounted more than the blue and the green actions whose returns are better known from the start and as more feedback is accumulated. This is similar to what is being assumed in the valuation equilibrium approach in which there is no ambiguity discount. Second, we find that the similarity-based reinforcement learning model explains the observed data better than the generalized Bayesian learning model.[3] We also observe that the patterns of choices stabilize well before the end of experiment and observe that the modal choices observed there correspond to those predicted by the valuation equilibrium.[4]

[3]The discount parameter estimated in the Bayesian model is very small too.

[4]We obtain a much better fit in terms of frequency of choices when considering the quantal valuation equilibrium, in which noisy best-responses are considered.

The rest of the paper makes the above claims precise. We start with a discussion of the related literature, we next provide a detailed description of the experimental design as well as of the various theories discussed above. Detailed statistical analysis is provided next. In the last part, we offer a general discussion, considering various robustness checks, an additional family of learning models this time based on imitation that we show does not explain our data well, and a discussion of how the long-run performance of subjects is ameliorated when the feedback for the red actions is made available state by state.

2 Related Literature

While the experimental literature on ambiguity is vast, only a few experimental papers look at ambiguous signals as we do (beyond Epstein and Halevy, we are only aware of Fryer et al (2019)). Note though that our experiment has a distinctive feature not present in the previous experiments on ambiguous signals. In our setting, the nature of the ambiguity of the received signals (feedback) is endogenously shaped by the choice of subjects. This endogenous character of the ambiguity has no counterpart in the previous experiments on ambiguity, as far as we know.

Our paper is related to other strands of literature beyond the references already mentioned. A first line of research related to our study is the framework of case-based decision theory as axiomatized by Gilboa and Schmeidler (1995). Compared to case-based decision theory, in the valuation equilibrium approach, the similarity weights given to the various actions in the various states happen to be endogenously shaped by the strategy used by the subjects, an equilibrium feature that is absent from the subjective perspective adopted in Gilboa and Schmeidler.

Another line of research related to our study includes the possibility that the strategy used by subjects would not distinguish behaviors across different states (Samuelson (2001), Mengel (2012) for theory papers and Grimm and Mengel (2012), Cason et al (2012) or Cownden et al. (2018) for experiments). Our study differs from that line of research in that subjects do adjust their behavior to the state but somehow mix the payoff consequences of some actions (the unfamiliar ones) obtained over different states, thereby revealing that our approach cannot be captured by a restriction on the strategy space, as arising in the literature just mentioned.

The analogy-based expectation equilibrium (Jehiel (2005) and Jehiel and Koessler (2008)) in which beliefs about other players' behaviors are aggregated over different states is also related to our study. One difference is that we are consider decision problems and not games. Yet, viewing nature as a player would allow us to see closer connections between the two approaches. To the best of our knowledge, no experiment in the vein of the analogy-based expectation equilibrium has considered environments similar to the one considered here.

The literature on selection neglect has also some connections with our study. On the

theory side, that literature includes the behavioral equilibrium introduced by Esponda (2008) in the context of adverse selection markets or Jehiel (2018) which develops an equilibrium model of selection neglect in an investment decision context in the vein of the analogy-based expectation equilibrium. On the experimental side, that literature includes Esponda and Vespa (2018), Enke (2019) or Barron et al. (2019), which conclude in various applications that subjects tend to ignore that the data they see are selected. In our setting, the data related to the red actions are selected to the extent that they are influenced by the strategy followed by agents (see the discussion above on the endogeneous weighting), and one can argue that subjects by behaving in agreement with the (generalized) valuation equilibrium do not seem to account for selection. Thus, our experimental setting can be viewed as providing an additional illustration of selection neglect in a novel context.

Another related recent strand of experimental literature is concerned with the failure of contingent reasoning and/or some form of correlation neglect (see Enke and Zimmerman (2019), Martinez-Marquina et al (2019) or Esponda and Vespa (2019)). Some of these papers (see in particular Martinez-Marquina et al.) conclude that hypothetical thinking is more likely to fail in the presence of uncertainty, which agrees with our finding that in the presence of aggregate feedback, subjects find it hard to disentangle the value of choosing the red action in the two states.

From another perspective, there is a number of contributions comparing reinforcement learning models to belief-based learning models in normal form games. While some of these contributions conclude that reinforcement learning models explain better the observed experimental data than belief-based learning models (Roth and Erev 1998, Camerer and Ho 1999), others suggest that it is not so easy to cleanly disentangle between these models (Salmon 2001, Hopkins 2002, Wilcox 2006). Our study is not much related to this debate to the extent that we consider decision problems and not games and that subjects do not immediately experience the payoff consequences of their choices (the feedback received concerns all subjects in the lab and subjects are only informed at the end how much they themselves earned). Relatedly the feedback received about some possible choices is aggregated over different states, which was not considered in the previous experimental literature. Despite these differences, in our context, relating Bayesian learning models to belief-based learning models suggest that these perform less well than their reinforcement learning counterpart, as in a number of these other works.

Finally, one should mention the experimental work of Charness and Levin (2005) who consider decision problems in which, after seeing a realization of payoff in one urn, subjects have to decide whether or not to switch their choices of urns. In an environment in which subjects have a probabilistic knowledge about how payoffs are distributed across choices and states (but have to infer the state from initial information), Charness and Levin observe that when there is a conflict between Bayesian updating and Reinforcement learning, there are significant deviations from optimal choices. While the conclusion that subjects may rely on reinforcement learning more than on Bayesian reasoning is somehow common in their study and our experiment, the absence of ex-ante statistical knowledge about the

distribution of payoffs across states in our experiment makes it clearly distinct from their experiment. In our view, the absence of ex-ante statistical knowledge fits the motivating economic examples mentioned in the introduction better.

3 Experimental design

Before describing the various theories put to a test, it is useful to describe more precisely the experimental setting. There are two states, $s = 1, 2$. In each state, the decision maker has to choose between two urns identified with a color, Blue and Red in state $s = 1$, Green and Red in state $s = 2$ where the Red urns have different payoff implications in states $s = 1$ and 2.[5] Each urn is composed of ten balls, black or white. When an urn is picked, one ball is drawn at random from this urn (and it is immediately replaced afterward). If a black ball is drawn this translates into a positive payment. If a white ball is drawn there is no payment. We conduct three treatments varying the composition of the Red urns across states and keeping the composition of Blue and Green fixed. One hundred initial draws are made for the Blue and Green urns with no payoff implication for participants, and all subjects are informed of the corresponding compositions of black and white balls drawn from these urns. Thus, as seen in Table 1, subjects have a precise initial view about the compositions of the Blue and Green urns (these urns correspond to the familiar choices in the motivating examples provided in the introduction). In the experiment, the Blue urn has 3 black balls out of ten and the Green urn has 7 black balls out of ten.

Blue	30 black (B)	70 white (W)
Green	68 black (B)	32 white (W)

Table 1 Information about the returns to the Blue and Green urns after 100 random draws as reported at the start of each session.

Concerning the Red urns, there is no initial information except that we make it clear that the compositions of these remains unchanged throughout the experiment. The Red urns correspond to the unfamiliar choices in the introductory examples. To guide their choices, subjects are provided with feedback about the compositions of the red urns as reflected by the colors of the balls that were previously drawn when a red urn either in state $s = 1$ or 2 was chosen. More precisely, there are twenty subjects and 70 rounds. In each round, ten subjects make a choice of an urn in state 1 and the other ten choose an urn in state 2. There are permutations of subjects between rounds so that every subject is in each state $s = 1$ or 2 the same proportion of the time. Between rounds, subjects receive feedback about the number of times the Green, Blue and Red urns were picked by

[5] In the instruction in Appendix A, Red is referred to as Red_1 in State 1 and Red_2 in State 2.

the various agents in the previous round, and for each color of the urn, they are informed of the number of black balls that were drawn. A typical feedback screen is shown in Figure 1.

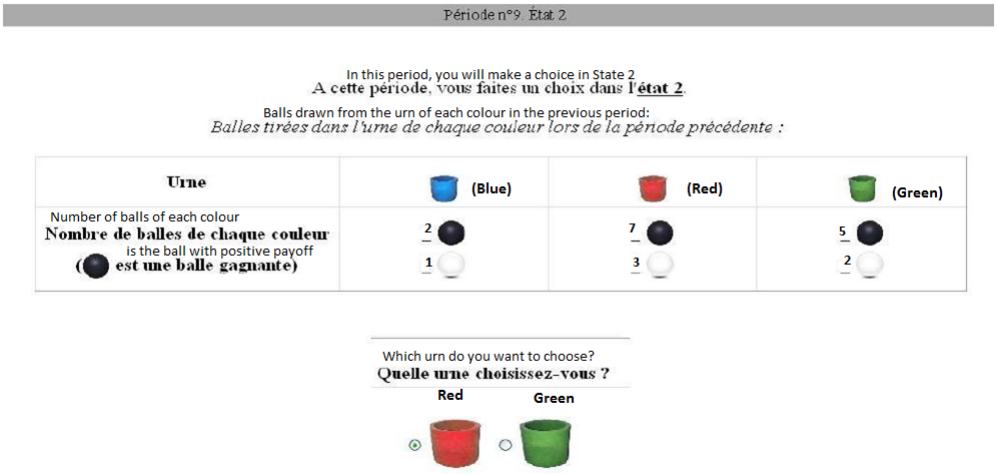


Figure 1 *Feedback structure for treatment sessions*

Note that in the case of the Red urns, this number aggregates the number of black balls drawn from both the Red urns picked in state $s = 1$ and the Red urns picked in state $s = 2$ mimicking the kind of coarse information suggested in the motivating examples. It should be highlighted that subjects were explicitly told that the compositions of the Red urn in state $s = 1$ and in state $s = 2$ need not be the same.

We consider three treatments T1, T2, T3 that differ in the composition of the Red urns as depicted in Figure 2, but note that we maintain the compositions of the Blue and Green urns in all treatments. The initial conditions in these various treatments are thus identical and any difference of behaviors observed in later periods can be safely attributed to the difference in the feedback received by the subjects across the treatments. In treatment 1, the best decision in both states $s = 1$ and 2 would require the choice of the Red urn. In treatment 2, the best decision would require picking the Red urn in state 1 but not in state 2. Finally, in treatment 3, the best decision would require picking the Red urn in state 2 but not in state 1.

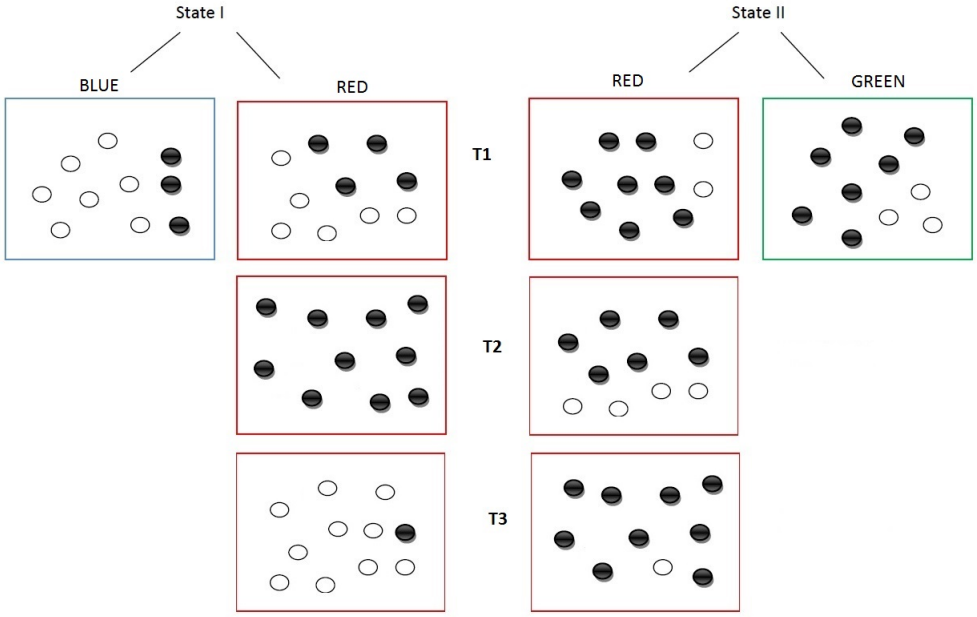


Figure 2 *Set up of the different treatment sessions*

4 Background and theory

In the context of our experiment, this section defines a generalization of the valuation equilibrium allowing for noisy best-responses in the vein of the quantal response equilibrium (McKelvey and Palfrey, 1995). We next propose two families of learning models, a similarity-based reinforcement learning model (allowing for coarse feedback on some alternatives, an ambiguity discount attached to those, and noisy best-responses) as well as a generalized Bayesian model (also allowing for noisy best-responses and a discount on alternatives associated to coarse feedback). The learning models will then be estimated and compared in terms of fit in light of our experimental data.

4.1 Quantal valuation equilibrium

In the context of our experiment, there are two states $s = 1$ and 2 that are equally likely. In state $s = 1$, the choice is between $Blue$ and Red_1. In state $s = 2$, the choice is between $Green$ and Red_2. The payoffs attached to these four alternatives are denoted by $v_{Blue} = 0.3$, v_{Red_1}, v_{Red_2} and $v_{Green} = 0.7$ where v_{Red_1} and v_{Red_2} are left as free variables to accommodate the payoff specifications of the various treatments.

A strategy for the decision maker can be described as $\sigma = (p_1, p_2)$ where p_i denotes

the probability that Red_i is picked in state $s = i$ for $i = 1, 2$. Following the spirit of the valuation equilibrium (Jehiel and Samet, 2007), a single valuation is attached to Red_1, Red_2 so as to reflect that subjects in the experiment only receive aggregate feedback about the payoff obtained when a Red urn is picked either in state $s = 1$ or 2. Accordingly, let $v(Red)$ be the valuation attached to Red. Similarly, we denote by $v(Blue)$ and $v(Green)$ the valuations attached to the Blue and Green urns, respectively.

In equilibrium, we require that the valuations are consistent with the empirical observations as dictated by the equilibrium strategy $\sigma = (p_1, p_2)$. This obviously implies that $v(Blue) = v_{Blue}$, $v(Green) = v_{Green}$ and less straightforwardly that

$$v(Red) = \frac{p_1 \times v_{Red_1} + p_2 \times v_{Red_2}}{p_1 + p_2} \tag{1}$$

whenever $p_1 + p_2 > 0$. That is, $v(Red)$ is a weighted average of v_{Red_1} and v_{Red_2} where the relative weight given to v_{Red_1} is $p_1/(p_1 + p_2)$ given that the two states $s = 1$ and 2 are equally likely and Red_i is picked with probability p_i for $i = 1, 2$.

Based on the valuations $v(Red)$, $v(Blue)$ and $v(Green)$, the decision maker is viewed as picking a noisy best-response where we consider the familiar logit parameterization (with coefficient λ). Formally,

Definition: A strategy $\sigma = (p_1, p_2)$ is a quantal valuation equilibrium if there exists a valuation system $(v(Blue), v(Green), v(Red))$ where $v(Blue) = 0.3$, $v(Green) = 0.7$, $v(Red)$ satisfies (1), and

$$\begin{aligned} p_1 &= \frac{e^{\lambda v(Red)}}{e^{\lambda v(Red)} + e^{\lambda v(Blue)}} \\ p_2 &= \frac{e^{\lambda v(Red)}}{e^{\lambda v(Red)} + e^{\lambda v(Green)}} \end{aligned}$$

It should be stressed that the determination of $v(Red)$, p_1 and p_2 are the results of a fixed point as the strategy $\sigma = (p_1, p_2)$ affects $v(Red)$ through (1) and $v(Red)$ determines the strategy $\sigma = (p_1, p_2)$ through the two equations just written.

We now briefly review how the quantal valuation equilibria look like in the payoff specifications corresponding to the various treatments. In this review, we consider the limiting case in which λ goes to ∞ (thereby corresponding to the valuation equilibria as defined in Jehiel and Samet, 2007).

Treatment 1: $v_{Red_1} = 0.4$ and $v_{Red_2} = 0.8$

In this case, clearly $v(Red) > v(Blue) = 0.3$ (because $v(Red)$ is some convex combination between 0.4 and 0.8). Hence, the optimality of the strategy in state $s = 1$ requires that the Red urn is always picked in state $s = 1$ ($p_1 = 1$). Regarding state $s = 2$, even if Red_2 were picked with probability 1, the resulting $v(Red)$ that would satisfy (1) would

only be $\frac{0.4+0.8}{2} = 0.6$, which would lead the decision maker to pick the Green urn in state $s = 2$ given that $v(Green) = 0.7$. It follows that the only valuation equilibrium in this case requires that $p_2 = 0$ so that the Red urn is only picked in state $s = 1$ (despite the Red urns being payoff superior in both states $s = 1$ and 2). In this equilibrium, consistency (i.e., equation (1)) implies that $v(Blue) < v(Red) = 0.4 < v(Green)$.

Treatment 2: $v_{Red_1} = 1$, $v_{Red_2} = 0.6$

In this case too, $v(Red) > v(Blue) = 0.3$ (because any convex combination of 0.6 and 1 is larger than 0.3) and thus $p_1 = 1$. Given that $v_{Red_2} < v_{Red_1}$, this implies that the lowest possible valuation of *Red* corresponds to $\frac{1+0.6}{2} = 0.8$ (obtained when $p_2 = 1$). Given that this value is strictly larger than $v(Green) = 0.7$, we obtain that it must that $p_2 = 1$, thereby implying that the Red urns are picked in both states. Valuation equilibrium requires that $p_1 = p_2 = 1$ and consistency implies that $v(Blue) < v(Green) < v(Red) = 0.8$.

Treatment 3: $v_{Red_1} = 0.1$, $v_{Red_2} = 0.9$

In this case, we will show that the Red urns are picked neither in state 1 nor in state 2. To see this, assume by contradiction that the Red urn would (sometimes) be picked in at least one state. This should imply that $v(Red) \geq v(Blue)$ (as otherwise, the Red urns would never be picked neither in state $s = 1$ nor 2). If $v(Red) < v(Green)$, one should have that $p_2 = 0$, thereby implying by consistency that $v(Red) = v_{Red_1} = 0.1$. But, this would contradict $v(Red) \geq v(Blue) = 0.3$. If $v(Red) \geq v(Green)$, then $p_1 = 1$ (given that $v(Red) > v(Blue)$), and thus by consistency $v(Red)$ would be at most equal to $\frac{0.1+0.9}{2} = 0.5$ (obtained when $p_2 = 1$). Given that $v(Green) = 0.7 > 0.5$, we get a contradiction, thereby implying that no Red urn can be picked in a valuation equilibrium.

As explained above the value of $v(Red)$ in the valuation equilibrium varies from being below $v(Blue)$ in treatment 3 to being in between $v(Blue)$ and $v(Green)$ in treatment 1 to being above $v(Green)$ in treatment 2, thereby offering markedly different predictions according to the treatment in terms of long run choices. Allowing for noisy as opposed to exact best-responses still allows us to differentiate the behaviors across the treatments but in a less extreme form (clearly, if $\lambda = 0$ behaviors are random and follow the lottery $50 : 50$ in every state and every treatment, but for any $\lambda > 0$, behaviors are different across treatments).

For each of the treatments, Table 2 shows how the predictions of valuation equilibrium differ from the optimal choice. Table 2 will be used to compare the modal choices in the final rounds of the experiment in the various treatments.

	T1	T2	T3
Optimal	(R,R)	(R,G)	(B,R)
Valuation Eq.	(R,G)	(R,R)	(B,G)

Table 2 Long run predictions

4.2 Learning Models

We will consider two families of learning models to explain the choice data observed in the various treatments of the experiment: A similarity-based version of reinforcement learning model in which choices are made on the basis of the valuations attached to the various colors of urns and valuations are updated based on the observed feedback, and a Bayesian learning model in which subjects update their prior belief about the composition of the Red urns based on the feedback they receive. In each case, we will assume that subjects care only about their immediate payoff and do not consider the possible information content that explorations outside what maximizes their current payoff could bring. This is -we believe- justified to the extent that in the experiment there are twenty subjects making choices in parallel and that the feedback is anonymous making the informational value of the experimentation by a single subject rather small. We will elaborate on this at the end of the Section after the exposition of the learning models.

4.2.1 Similarity-based reinforcement learning

Standard reinforcement learning models assume that strategies are reinforced as a function of the payoff obtained from them. In the context of our experiment, subjects receive feedback about how the choices made by all subjects in the previous period translated into black (positive payoff) or white (null payoff) draws. More precisely, the feedback concerns the number[6] of Black balls drawn when a *Blue*, *Green* or *Red* urn was picked in the previous period as well as the number of times an urn with that color was then picked. Accordingly, at each time $t = 2, ...70$, one can define for each possible color $C = B, R, G$ (for Blue, Red, Green) of the urn(s) that was picked at least once at $t-1$:

$$UC_t = \frac{\#(Black\ balls \text{ drawn in urns with color } C \text{ at } t-1)}{\#(\text{an urn with color } C \text{ picked at } t-1)}. \tag{2}$$

UC_t represents the strength of urn(s) with color C as reflected by the feedback received at t about urns with such a color. Note the normalization by #(an urn with color C picked at $t-1$) so that UC_t is comparable to a single payoff attached to choosing an urn with color C.

We will let BC_t denote the value attached to an urn with color C at time t and BC_{init} denote the initial value attached to an urn with that color. For *Green* and *Blue* there is

[6]The symbol # is used to refer to number.

initial information and it is natural to assume that

$$\begin{aligned} BB_{init} &= \frac{30}{100} = 0.3 \\ BG_{init} &= \frac{68}{100} = 0.68 \end{aligned}$$

whereas for *Red*, the initial value BR_{init} is a priori unknown and it will be estimated in light of the observed choice data.

Dynamics of BC_t:
Concerning the evolution of BC_t, we assume that for some (ρ_U, ρ_F), we have:[7]

$$\begin{aligned} BR_t &= \rho_U \times BR_{t-1} + (1 - \rho_U) \times UR_t \\ BB_t &= \rho_F \times BB_{t-1} + (1 - \rho_F) \times UB_t \\ BG_t &= \rho_F \times BG_{t-1} + (1 - \rho_F) \times UG_t \end{aligned}$$

In other words, the value attached to color C at t is a convex combination between the value attached at $t-1$ and the strength of C as observed in the feedback at t. Observe that we allow the weight to be assigned to the feedback to be different for the Red urns on the one hand and the Blue and Green urns on the other to reflect the idea that when a choice is better known as is the case for more familiar alternatives (here identified with urns *Blue* and *Green*) the new feedback may be considered as less important to determine the value of it. Accordingly, we would expect that ρ_F is larger than ρ_U, and we will be concerned whether this is the case in our estimations.[8]

Choice Rule:
Given that the feedback concerning the Red urns is aggregated over states $s = 1$ and 2, there is extra ambiguity as to how well BR_t represents the valuation of Red_1 or Red_2 as compared to how well BG_t or BB_t represent the valuations of *Blue* and *Green*.

The valuation equilibrium (or its quantal extension as presented above) assumes that BR_t is used to assess the strength of Red_s whatever the state $s = 1, 2$. In line with the literature on ambiguity aversion as experimentally initiated by Ellsberg (1961), it is reasonable to assume that when assessing the urn Red_s, $s = 1, 2$, subjects apply a discount

[7] In case no urn of color C was picked at $t-1$, then $UC_t = BC_{t-1}$ so that $BC_t = BC_{t-1}$.

[8] Many variants could be considered. For example, one could have made the weight of the new feedback increase linearly or otherwise with the number of times an urn with that color was observed. One could also have considered that the weight on the feedback is a (decreasing) function of t to reflect that as more experience accumulates, new feedback becomes less important. These extensions did not seem to improve how well we could explain the data and therefore, we have chosen to adopt the simpler approach just described.

$\delta \geq 0$ to BR_t.[9] Allowing for noisy best-responses in the vein of the logit specification, this would lead to probabilities p_{1t} and p_{2t} of choosing Red_1 and Red_2 as given by

$$p_{1t} = \frac{e^{\lambda(BR_t - \delta)}}{e^{\lambda(BR_t - \delta)} + e^{\lambda BB_t}}$$
$$p_{2t} = \frac{e^{\lambda(BR_t - \delta)}}{e^{\lambda(BR_t - \delta)} + e^{\lambda BG_t}}$$

The learning model just described is parameterized by $(\rho_U, \rho_F, \delta, \lambda, BR_{init})$. In the next Section, these parameters will be estimated pooling the data across all three treatments (and using the maximum likelihood method). Particular attention will be devoted to whether $\delta > 0$ is needed to explain the data better, whether $\rho_F > \rho_U$ as common sense suggests, as well as to the estimated value of λ and the obtained likelihood for comparison with the Bayesian model to be described next.

Assuming the ambiguity discount parameter δ is 0 and considering the limit as the number of subjects tends to infinity, it is readily verified that the steady states of the similarity-based reinforcement learning model just described coincides with the quantal valuation equilibria defined in the previous subsection (as long as ρ_U and ρ_F are strictly less than 1).[10] Of course, a similar steady state could be defined when $\delta > 0$, but as seen from our estimation exercise, $\delta = 0$ fits best our experimental data and thus this extension will not be needed for our purpose. A remaining question is whether the learning dynamic converges and for which values of the model parameters as well as how the dynamic is affected when the number of subjects is finite as in our experiment. We do not provide a comprehensive analysis of this here, but we have run simulations for the specification ($\rho_U = 0.43$, $\rho_F = 0.599$, $\delta = 0, \lambda = 5.24$, $BR_{init} = 0.42$) that corresponds to the estimation of this model given our experimental data. We find that the long run frequencies of Red_1 and Red_2 are 75% and 20% in Treatment 1; 90% and 70% in Treatment 2; and 41% and 13% in Treatment 3. In Appendix D, we report the simulation and observe not much variation of these frequencies after round 70 (the duration of our experiment) onwards.

[9] One possible rationale following the theoretical construction of Gilboa and Schmeidler (1989) is that the proportion of Black balls in Red_1 and Red_2 is viewed as being in the range [BR - δ, BR + δ] and that subjects adopt a maxmin criterion, leading them to consistently use $BR - \delta$ to assess both Red_1 and Red_2. More elaborate specifications of ambiguity would be hard to estimate given the nature of our data.

[10] To see this, assume that p_{1t} and p_{2t} remain constant over time and equal to p_1 and p_2, respectively. By the law of large number, the values of UR_t, UB_t and UG_t should be arbitrarily close to $v(Red)$, $v(Blue)$ and $v(Green)$ as defined in the previous Section, and thus BR_t, BB_t and BG_t should also converge to these values. Putting this together with the observation that when BR_t, BB_t and BG_t remain constant and equal to $v(Red)$, $v(Blue)$ and $v(Green)$, respectively, the resulting values of p_{1t} and p_{2t} would remain constant and satisfy the conditions shown in the definition of the quantal valuation equilibrium yields the desired result.

4.2.2 Generalized Bayesian Learning Model

As an alternative learning model, subjects could form some initial prior belief regarding the compositions of Red_1 and Red_2, say about the chance that there are k_i black balls out of 10 in Red_i, and update these beliefs after seeing the feedback using Bayes' law.

Let us call $\beta_{init}(k_1, k_2)$ the initial prior belief of subjects that there are k_i black balls out of 10 in Red_i. In the estimations, we will allow the subjects to consider that the number of black balls in either of the two Red urns can vary between $k_{\inf}$ and $k_{\sup}$ with $0 \leq k_{\inf} \leq k_{\sup} \leq 10$ and we will consider the uniform distribution over the various possibilities.

With uniform distribution, for any $(k_1, k_2) \in [k_{\inf}, k_{\sup}]^2$

$$\beta_{init}(k_1, k_2) = \frac{1}{(k_{\sup} - k_{\inf} + 1)^2},$$

and $\beta_{init}(k_1, k_2) = 0$ otherwise. The values of $k_{\inf}$ and $k_{\sup}$ will be estimated.

Of course, the family of prior just considered is somewhat arbitrary. For robustness check, we have also considered another class of priors, in which the maximum prior probability is assigned to $k = 5$ and then the probability decreases linearly in a symmetric way as k moves away from 5 (see details below). Together with the uniform distribution, we believe such variations of the prior allow us to span a relatively large range, and as will turn out within such priors, the best fit with our experimental data is obtained for the uniform distribution, thereby explaining our choice to present the Bayesian learning model with uniform priors.

Dynamics of the beliefs:

To simplify the presentation a bit, we assume there is no learning on the urns *Blue* and *Green* for which there is substantial initial information. At time $t + 1$, the feedback received by a subject can then be formulated as (b, g, n) where b, g are the number of blue and green urns respectively that were picked at t, and n is the number of black balls drawn from the *Red* urns. In the robustness checks, we allow for Bayesian updating also on the compositions of the Blue and Green urns, and obtain that allowing for learning on those urns does not change our conclusion.

To further simplify the presentation, we assume that in the feedback subjects are exposed to, subjects assume there is an equal number of states $s = 1$ and $s = 2$ decisions (allowing the subjects to treat these numbers as resulting from a Bernoulli distribution would not alter our conclusions, see the robustness check section for elaborations). In this case, the feedback can be presented in a simpler way, because knowing (b, g, n) now allows subjects to infer that $m_1 = 10 - b$ choices of *Red* urns come from state $s = 1$ and $m_2 = 10 - g$ choices of *Red* urns come from state $s = 2$. Accordingly, we represent the

feedback as (m_1, m_2, n) where m_i represents the number of Red_i that were picked. Clearly, the probability of observing m_1, m_2, n when there are k_1 and k_2 black balls in Red_1 and Red_2 respectively is given by:

$$Pr(m_1, m_2, n|k_1, k_2) = \sum_{\substack{n_1 \leq m_1 \\ n_2 \leq m_2 \\ n_1+n_2=n}} \left(\begin{array}{c} m_1 \\ n_1 \end{array} \right) \left(\begin{array}{c} m_2 \\ n_2 \end{array} \right) (k_1/10)^{n_1}(1-k_1/10)^{m_1-n_1}(k_2/10)^{n_2}(1-k_2/10)^{m_2-n_2}$$

where $\left(\begin{array}{c} a \\ b \end{array} \right) = \frac{a!}{(a-b)!b!}$ for integers a, b with $a \geq b$.

The posterior at $t+1$ about the probability that there are k_1 and k_2 black balls out of ten in Red_1 and Red_2 after observing (m_1, m_2, n) at t is then derived from Bayes' law by

$$\beta_{t+1}(k_1, k_2) = \frac{\beta_t(k_1, k_2) \cdot \Pr(m_1, m_2, n|k_1, k_2)}{\sum_{r_1, r_2} \beta_t(r_1, r_2) \cdot \Pr(m_1, m_2, n|r_1, r_2)}.$$

with $\beta_1(k_1, k_2) = \beta_{init}(k_1, k_2)$.

Define $v_t^{Bayes}(Red_i) = \sum_{k_i} \frac{k_i}{10} \beta_t(k_i)$ where $\beta_t(k_i) = \sum_{k_{-i}} \beta_t(k_i, k_{-i})$ as the time t expected proportion of black balls in Red_i given the distribution β_t.

Choice Rule:

As for the similarity-based reinforcement learning model, we allow for noisy best responses and introduce an ambiguity discount δ for the evaluation of the Red urns.[11] Accordingly, the probabilities p_{1t} and p_{2t} of choosing Red_1 and Red_2 at time t in the generalized Bayesian learning model are given by:

$$p_{1t} = \frac{e^{\lambda(v_t^{Bayes}(Red_1)-\delta)}}{e^{\lambda(v_t^{Bayes}(Red_1)-\delta)} + e^{\lambda v(Blue)}}$$
$$p_{2t} = \frac{e^{\lambda(v_t^{Bayes}(Red_2)-\delta)}}{e^{\lambda(v_t^{Bayes}(Red_2)-\delta)} + e^{\lambda v(Green)}}$$

where as our simplification implies we assume that $v(Blue) = 0.3$ and $v(Green) = 0.7$.[12]

Studying the dynamics of the above Bayesian learning model is a bit cumbersome for general specifications of $(k_{\inf}, k_{\sup}, \lambda, \delta)$. To illustrate how it leads to predictions markedly

[11] Some might dispute that the ambiguity discount is not so much in the spirit of the Bayesian model in which case one should freeze this parameter to be 0. We adopt a more permissive view about the Bayesian learning model and regard δ as a parameter to be estimated.

[12] As previously mentioned, we present a model that allows subjects to update $v(Blue)$ and $v(Green)$ according to Bayes rule in the robustness checks.

different from those of the valuation equilibrium, consider the extreme case in which $\delta = 0$, $k_{\text{inf}} = 0$, $k_{\text{sup}} = 10$ and $\lambda = \infty$ assuming the number of subjects is arbitrarily large. Then in all treatments, Red_2 is not chosen to start with given that it is perceived to deliver 0.5 in expectation, which is less than 0.7. As a result, subjects can safely attribute the feedback they receive about Red to be coming from Red_1. This in turn implies (considering the limiting case with infinitely large populations of subjects) that subjects perfectly learn the value of Red_1 and learn nothing about the value of Red_2. Thus, in all subsequent periods, subjects play Red_1 in treatments 1 and 2 and never play Red_1 in treatment 3, and they never play Red_2 in any of the treatments (by contrast, Red_2 was played in treatment 2 in the valuation equilibrium).[13] To get a sense of the predictions of the Bayesian learning model for more general specifications of the parameters, we have run simulations for the case in which $\delta = 0$, $k_{\text{inf}} = 3$, $k_{\text{sup}} = 7$ and $\lambda = 7.5$, which correspond to our estimation given the experimental data. We find that after 70 rounds the proportions of Red_1 and Red_2 are 80% and 30% in Treatment 1; 90% and 52% in Treatment 2; and 47% and 12% on Treatment 3. The most notable difference with respect to the simulation for the generalized reinforcement learning concerns the share of Red_2 choices in Treatment 2, which switches from 70% to 52%.

Our Bayesian learning model has some simplifying assumptions. In particular, it does not incorporate elements of experimentation that would prevail in traditional multi-arm-bandit problems. To the extent that some Red choices are made with positive probability at least in one state, we believe the conditions of the feedback that we consider would make experimentation of very little value, and in fact of no value at all when the pool gets arbitrarily large. To get a sense of this, assume only Red_1 is (sometimes) chosen in the pool. If a subject decides to experiment by choosing Red_2, she/he will know that one instance of the Red actions in the feedback corresponds to Red_2, but the feedback she/he receives will not inform her/him much about the return to Red_2 as most of the return observations would correspond to Red_1 and the subject would have no way to learn much (anything, in the limit of a very large pool) about the value of Red_2. The only case in which experimentation could be of significant value is if the Red action were rarely chosen, but this does not arise in our experimental data. Of course, from a theoretical viewpoint, understanding the value of experimentation better in a setting with coarse feedback would be interesting for smaller sizes of the pool. However, we suspect the derived effect would not be very significant for our experiment in which the pool -consisting of twenty subjects- was not so small.

[13] The obtained behaviors can also be interpreted as a self-confirming equilibrium in which the theory about the value of Red_1 would rationalize not experimenting with this choice and despite the coarseness of the feedback about the Red urns, subjects would be able to perfectly infer the value of Red_2.

5 Results

5.1 Further Description of the Experimental Design

The computerized experiments were conducted in the Laboratory at Maison de Sciences Economiques (MSE) between March 2015 and November 2016, with some additional sessions running in March 2017. Upon arrival at the lab, subjects sat down at a computer terminal to start the experiment. Instructions were handed out and read aloud before the start of each session.

The experiment consisted of three main treatments which varied in the payoffs of the Red urns as explained above. In addition, we had two other treatments referred to as controls in which subjects received state-specific feedback about the Red urns, i.e the feedbacks for Red_1 and Red_2 appeared now in two different columns, for the two payoff specifications of treatments 1 and 2. The purpose of these control treatments was to check whether convergence to optimal choices was observed in such more standard feedback scenarios.

Each session involved 18-20 subjects[14] and four sessions were run for each treatment and control. Overall, 235 subjects drawn from the participant pool at the MSE -who were mostly students- participated in the experiment. Each session had seventy rounds.

In all treatments, all sessions, and all rounds, subjects were split up equally into two states, State 1 and State 2. Subjects were randomly assigned to a new state at the start of each round. The subjects knew the state they were assigned to but did not know the payoff attached to the available actions in each state.[15] In each state, players were asked to choose between two actions as detailed in Figure 1. The feedback structure for the main treatments was as explained above. For the control group, the information structure was disaggregated. We use this as a baseline to show that under a simpler feedback structure, individuals learn optimally the best available option.

Subjects were paid a show-up fee of 5 €. In addition to this, they were given the opportunity to earn 10 € depending on their choice in the experiment. Specifically, for each subject, two rounds were drawn at random and a subject earned an extra 5 € for each black ball that was drawn from their chosen urn in these two rounds. The average payment was around 11 € per subject, including the turn-up fee. All of the sessions lasted between 1 hour and 1.5 hours, and subjects took longer to consider their choices at the start of the experiment.

5.2 Results

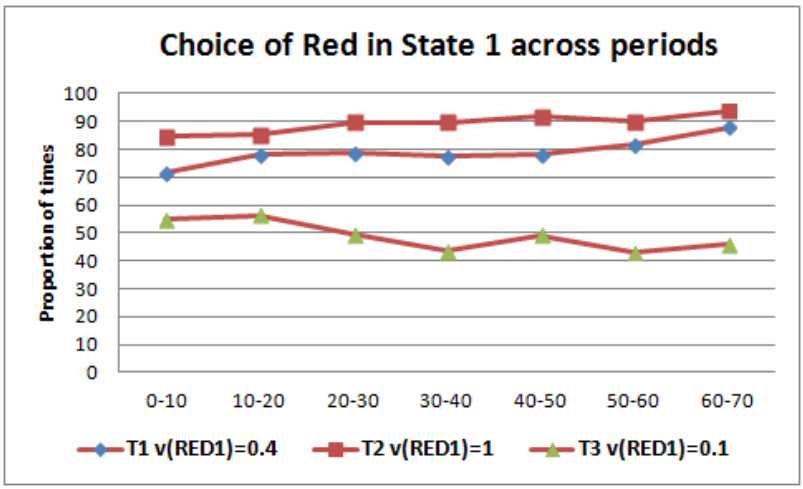
We first present descriptive statistics and next present the structural analysis.

[14] Note that when 18 subjects participated in the session, Bayes updating was modified accordingly.

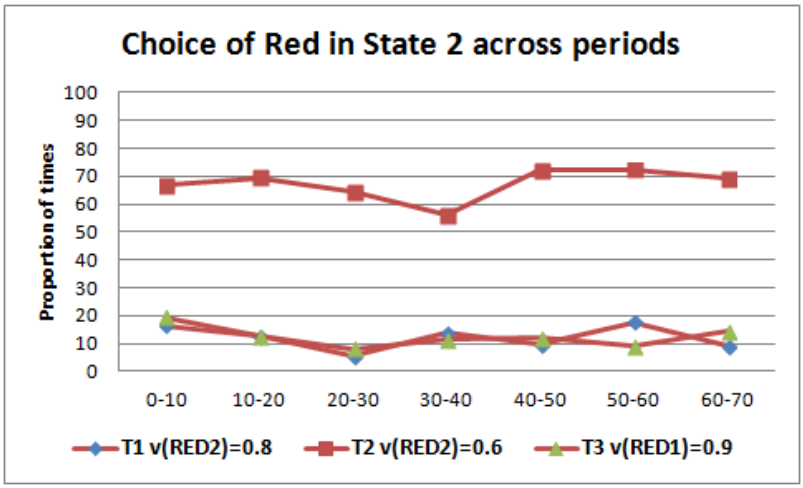
[15] For urns *Blue* and *Green*, they had initial information, as explained in the Introduction.

5.2.1 Preliminary findings

In Figure 3, we report how the choices of urns vary with time and across treatments. Across all these sessions, initially, subjects are more likely to choose the Red urn than the Blue urn in state 1 and they are more likely to choose the Green urn than the Red urn in state 2. This is, of course, consistent with most theoretical approaches including the ones discussed above given that the Green urn is more rewarding than the Blue urn and the Red urns look (at least initially) alike in states 1 and 2.



(a) Red1 across treatments



(b) Red2 across treatments

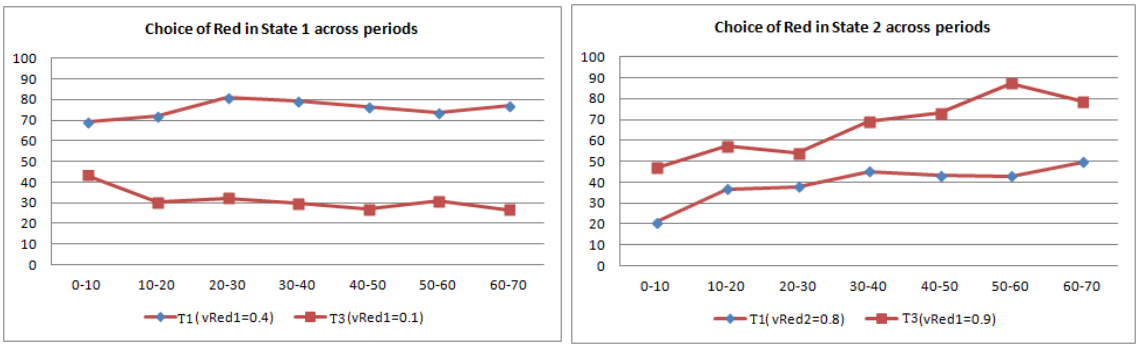
Figure 3 Evolution of choice across treatments with aggregated feedback

The more interesting question concerns the evolution of choices. Roughly, in state 1, we see toward the final rounds, a largely dominant choice of the Red urns in treatments 1 and 2 whereas Red in state 1 is chosen less than half the time in treatment 3. The modal choices observed in the final rounds agree for state 1 with the predictions of valuation equilibrium (which coincide with the optimal choices).

Concerning state 2, we see that in the final rounds, the Red urns are rarely chosen in treatments 1 and 3, and are frequently chosen in treatment 2. This agrees with the prediction of valuation equilibrium (and stands in sharp contrast with the optimal choices).

Overall, the qualitative differences of the choices in the final rounds among the three treatments and the two states are in line with the prediction of the valuation equilibrium even if some noise in the best-response is needed especially for treatment 3 in state 1 to explain why about 40% of choices correspond to Red.[16]

In Figure 4, with state-specific feedback for the Red urns, we see a clear trend toward the optimal choices even if some noise would be needed to explain why only 49% of choices correspond to Red in state 2 in Control 1. In contrast to the feedback structure in the treatment group, we see that disaggregating feedback on the Red urns across states, players learn the optimal choice. In line with section 3.1, the fine feedback helps the agent attach a valuation $v(Red_1)$ and $v(Red_2)$ separately for the *Red* urns in the two states instead of a joint valuation $v(Red)$. Due to this finer feedback structure, the simple heuristic of reinforcement learning leads to an optimal choice, unlike in the main treatments in which an analogous reinforcement learning heuristic leads to valuation equilibrium.



(a) Red1 in control

(b) Red2 in control

Figure 4 Evolution of choice across treatments with dis-aggregated feedback

[16]We note that the large share of Red chosen in state 2 of treatment 2 is not in line with the noisy version of the Bayesian learning model as explained above.

5.2.2 Statistical estimations

Similarity-based reinforcement learning

The estimations of the parameters of the similarity-based reinforcement learning model together with the corresponding log likelihood[17] are given in the following Table.

Table 3 Parameters for similarity-based reinforcement learning model

ρ_U	ρ_F	δ	BR_{ini}	λ	L
0.43	0.599	0.00	0.42	5.24	7626.6
[0.4, 0.49]	[0.55,0.64]	[0, 0.0009]	[0.38, 0.49]	[5.04, 5.39]	-

Note: Confidence interval at 95% are reported in brackets for the restricted estimators. (See Ketz 2018 for details).

Concerning the likelihood, by way of comparison, a completely random choice model wherein every state, subjects would randomize 50:50 between the two choices would result in a negative log likelihood of L=11402, which is much higher than 7626.6. More generally, the similarity-based reinforcement learning model explains data much better than any model in which behavior would not be responsive to feedback. [18]

We now discuss the most salient aspects of the estimations.

The finding that $\rho_F > \rho_U$ seems natural as mentioned above, to the extent that for the familiar urns, the feedback should affect less how the valuations are updated.

The finding that BR_{init} is slightly below 0.5 may be interpreted along the following lines. In the absence of any information, an initial value of 0.5 would be the one dictated by the principle of insufficient reason, but the uncertainty attached to the unfamiliar urns may lead to some extra discount in agreement with some form of ambiguity aversion as reported in Ellsberg.[19]

The most interesting observation concerns δ which is estimated to be 0. Even though the feedback for the Red urns is ambiguous (because it is aggregated over the two states), the valuations for Red are not discounted as if subjects were ambiguous neutral from that perspective.

Thus, what our estimation suggests is that while there may be some (mild) initial ambiguity aversion relative to the unfamiliar choices (as reflected by BR_{init} being smaller than 0.5), no ambiguity discount seems to be applied to the valuation of Red despite the ambiguity attached to the feedback received about the Red urns.

[17] Likelihood throughout the paper refers to the negative of the log likelihood. Thus, the lower the likelihood, the better the model (See textbook Train 2003 for further details). Standard errors are reported in brackets.

[18] Optimizing on the probability of Red_1 vs Red_2 in such a model, would lead to assume that Red_1 is chosen with probability p_1=0.7 and Red_2 is chosen with probability p_2= 0.3 with a negative log likelihood of L=9899.6 which is much higher than 7626.6

[19] The difference $0.5 - 0.44 = 0.06$ can be interpreted as measuring the ambiguity aversion of choosing an unfamiliar urn when no feedback is available.

Generalized Bayesian learning model:

The estimated parameters for the generalized Bayesian learning model are given in the following Table.

Table 4 Parameter for Bayesian model with bounds

λ	k_{inf}	k_{sup}	δ	L
7.488	3	7	0.003	8816.2
[7.43, 7.52]	-	-	[0.001, 0.005]	-

The value of δ =0.003 implies that with the Bayesian model, the subjects show some mild form of ambiguity aversion. However we cannot statistically reject the hypothesis that $\delta = 0$, which implies that with the Bayesian model too, there is no significant ambiguity discount similar to what we found in the similarity-based reinforcement learning estimations. For the support of initial prior, we found that $k_{\text{inf}} = 3$ and $k_{\text{sup}} = 7$.[20] We also note that the value of λ is higher than that for the reinforcement model.

Comparing the two models:

Maybe the most important question is which of the Bayesian learning model or the reinforcement learning model explains the experimental data best. We use three methods of comparison, all establishing that the reinforcement learning model outperforms the Bayesian learning model. First, looking at the likelihood of the two models, we see that the Bayesian learning model explains the data less well than the similarity-based reinforcement learning model. Second, to account for the difference in the number of parameters between the two models, we use the Bayesian Information Criterion (BIC) or Schwarz criterion (also SBC, SBIC). BIC is a criterion for model selection among a finite set of models where the model with the lower BIC is closer to the data generating process. It is based, in part, on the likelihood function to determine the goodness of fit in the two models accounting for the number of parameters, formally defined as

$$\text{BIC} = \ln(n)k - 2\ln(L^*)$$

where L^* is value of maximized likelihood of model M, n is the number of observations, k is the number of parameters estimated by the model. As seen from table 5, we can conclude that the reinforcement model performs better than the Bayesian one in explaining the data.

Finally, we perform a Vuong test[21] to compare the performance of the two models statistically. Under the null hypothesis H_0 , that both models perform equally well, we

[20]The value of the bounds correspond to v_{Blue}=0.3 and v_{Green}=0.7 respectively and so one may speculate that maybe the compositions of the familiar urns serve as anchoring the support of the priors. Observe that because best-responses are noisy, the derived support does not imply that the Red urn is always picked in state 1 and never picked in state 2.

[21]See Merkel et al. 2016 for more details.

Table 5 BIC values for the two competing models

Valuation model	Bayesian model
1.52 x 10^4	1.763 x 10^4

conclude that the null can be rejected in favor of the reinforcement model. Specifically,

$$H_0 = E(L(\theta_R; x_d)) = E(L(\theta_B; x_d))$$

$$H_a = E(L(\theta_R; x_d)) \neq E(L(\theta_B; x_d))$$

where x_d is the collection of observed individual data points, θ_R is the set of parameters estimated via reinforcement learning, θ_B is the set of parameters estimated via Bayesian learning, $L(\theta_R; x_d)$ is the log likelihood under reinforcement model and $L(\theta_B; x_d)$ is the log likelihood under Bayesian learning model for each data point d. The Vuong statistics is then defined by

$$V_{stat} = \sqrt{N}\frac{\bar{m}}{S_m}$$

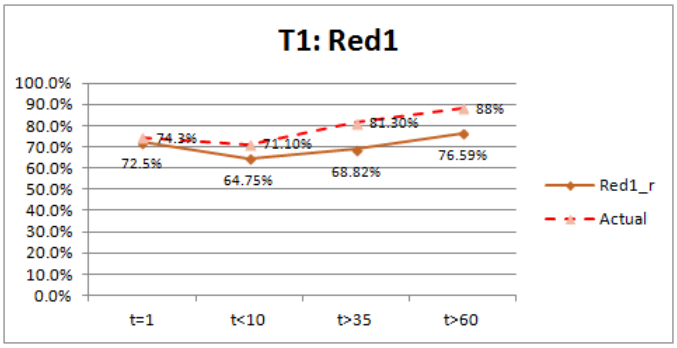
where $\bar{m} = E(L(\theta_R; x_d)) - E(L(\theta_B; x_d))$ for each individual d, N is the total number of observations and S_m is the sample standard deviation.

V_{stat} tests the null hypothesis (H_0) that the two models are equally close to the true data generating process, against the alternative H_1 that one of the two model is closer. [22] The obtained $V_{stat} = 25.01$ being large and positive implies that the reinforcement model is a better fit to our experimental data than the Bayesian learning model. This is in line with the findings derived with the BIC methodology.

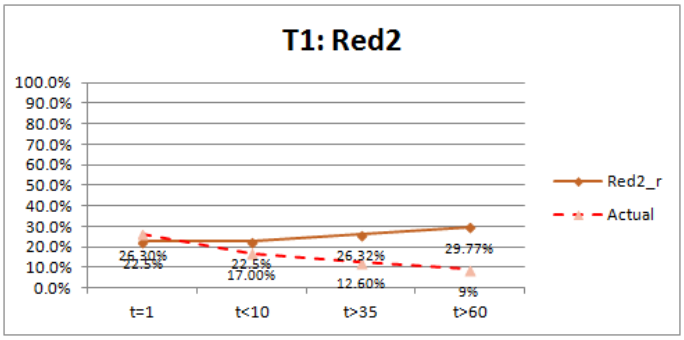
5.3 Comparing the Reinforcement learning model to the data

It is of interest to see how the obtained frequencies of choices as generated by the similarity-based reinforcement learning model with estimations, as reported in Table 3, compared to the observed frequencies from our experimental data. In Figure 5, we report the simulated frequencies of urn choices using the reinforcement model across all periods and treatments. Across all these sessions, our simulated frequencies remain close to the actual frequencies with a slightly less good fit in Treatment 1. Allowing for a different λ in Treatment 1, we observe that a larger lambda significantly improves the fit in this treatment as shown in Appendix C.

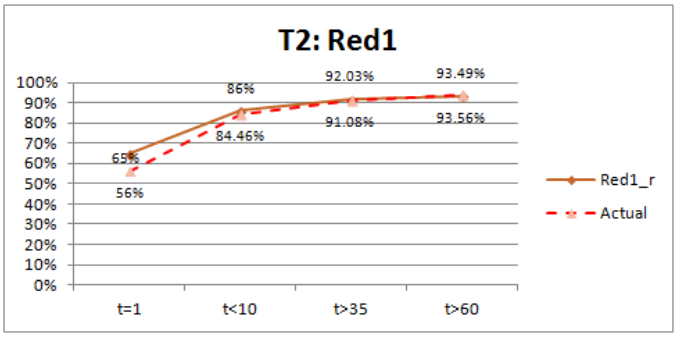
[22] Vuong test compares the predicted probabilities of two non nested models. It computes the difference in likelihood for each observation i in the data. A high positive V_{stat} implies Model 1 is better than Model 2 where $\bar{m}$=log($Pr(x_i|Model1)$-log($Pr(x_i|Model2)$



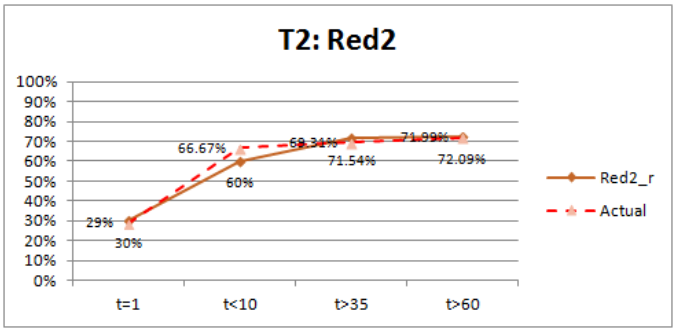
(a) Treatment 1: State 1



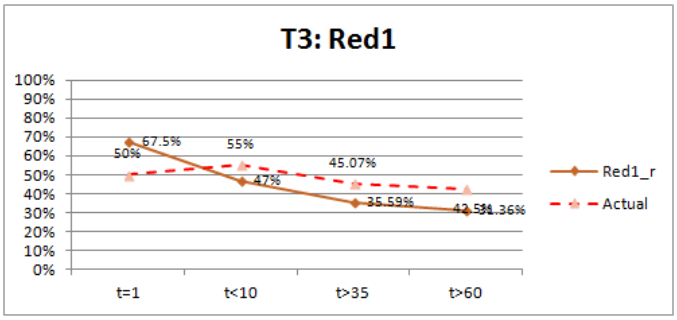
(b) Treatment 1: State 2



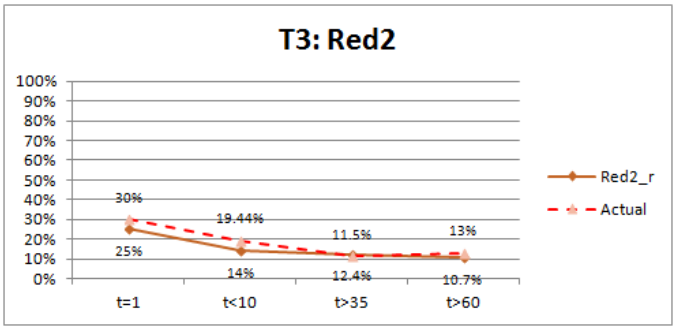
(c) Treatment 2: State 1



(d) Treatment 2: State 2



(e) Treatment 3: State 1



(f) Treatment 3: State 2

Figure 5 Simulated choices using the reinforcement model

5.4 Categorization of individuals

While we have established that the similarity-based reinforcement learning model explains the data better than its Bayesian counterpart, it is of interest to see how the fit as generated by such a model with estimations as reported in Section 5.2.2 vary across individuals over the various rounds. For each round, we compute the log likelihood of observing the choices

by the individual, and we then add up for this individual the log likelihood across the 70 rounds so to determine which model fits the overall behavior of the individual better.

That is, for each individual, we ask which of the two learning models, the generalized reinforcement learning model with parameters as in Table 3 or the generalized Bayesian model with parameters as in Table 4 explains the observed choices across all rounds better. For the model comparison, we use the Vuong test as explained above. Overall, we find that 82.97% of players can be categorized as reinforcement learners and the rest as Bayesian learners.

Comparing across treatments, we find that T1 has the highest proportion of Reinforcement learners (93.5%). In T2, the proportion of similarity-based reinforcement learners is 79.22% followed by 76.25% in T3. The higher proportion of Reinforcement learners in T1 can be related to the previous observation that in aggregate the Reinforcement learning model with a lower level of noise allows for a better fit in T1.

The pool of 235 subjects had very similar backgrounds across treatments, making it impossible to relate the categorization in learning types to the background.We have also compared the overall payoff as measured by the sum of black balls obtained over the 70 rounds across the population of subjects categorized as reinforcement learners or Bayesian learners. For reinforcement learners, the average was 42.09 black balls whereas, for Bayesian learners, an average of 40.55 black balls were observed. But, this difference is not statistically significant.

6 Discussion

In this Section, we consider several variants of the basic learning models introduced so far. The study of these variants can be seen as providing robustness checks for our main insights. We then discuss an alternative model based on imitation that we think might be the result of another possible heuristic used by subjects in the context of our experiment, but such a model turns out not to explain well our data. Finally, we discuss the payoff performances obtained by subjects in the main conditions of our experiment with those obtained when subjects receive fine feedback.

6.1 Robustness Checks

As many variants of reinforcement learning models and Bayesian models could be considered, we review a few of these here and suggest that our basic conclusions remain the same in these variants. In each case, the reported estimation relies on the same methodology as in Section 5.

Similarity-based reinforcement learning model

Regarding reinforcement models, we consider the following variants. First, we allow the speed of adjustment of the valuation of the Red urns to differ across the two states as a large imbalance in the number of green urns as opposed to blue urns in the feedback is indicative that the feedback concerned more Red urns in state 1 than in state 2.[23] Specifically, we now introduce a new parameter μ and specify the weight on the previous valuation to satisfy

$$\rho_{U1} = \rho_U \times [1 - \mu \cdot (\frac{NB}{NG + NB} - 0.5)]$$
$$\rho_{U2} = \rho_U \times [1 - \mu \cdot (\frac{NG}{NG + NB} - 0.5)]$$

where NB and NG are the respective numbers of Blue and Green urns appearing in the feedback. One would expect μ to be negative so that when NB is observed to be smaller than NG, subjects infer the new feedback on the Red urns is more informative on the composition of Red in state 1 than Red in state 2 and thus update more the valuation of Red_1 than Red_2. The estimations of this extended model are reported in the following table.

Table 6 Parameters for variant1 of similarity-based reinforcement learning model

ρ_U	ρ_F	δ	BR_{ini}	λ	μ	L
0.45	0.6	0.00	0.45	5.2	0.108	7626.3
[0.40, 0.49]	[0.55,0.64]	[0, 0.001]	[0.39, 0.5]	[5.02, 5.37]	[-0.26, 0.48]	-

Our estimation yields $\mu > 0$ in contrast to what might have been expected, but note that it is not significant and that $\mu = 0$ cannot be rejected. Thus, this extended model does not explain the data better than the simpler version considered above.

A different idea somewhat related to the one just discussed is that subjects would apply a different discount to the Red urn in state 1 and 2 maybe because they would consider the feedback for the Red urns to be more indicative of Red in state 1 than in state 2 (again maybe because of the imbalance of the number of Blue and Green urns in the feedback). This leads us to consider an extended version with two different discounts δ_1 and δ_2 for Red in state 1 and 2 while keeping the other aspects of the dynamics unchanged as compared to the main reinforcement learning model. That is, the only change in this variant is in the choice rule with now two discount parameters δ_1 and δ_2.

[23] This is in some sense making use of some qualitative features of the Bayesian model to improve the reinforcement learning model.

Choice Rule:

$$p_{1t} = \frac{exp^{\lambda(BR_t - \delta_1)}}{exp^{\lambda(BR_t - \delta_1)} + exp^{\lambda BB_t}}$$

$$p_{2t} = \frac{exp^{\lambda(BR_t - \delta_2)}}{exp^{\lambda(BR_t - \delta_2)} + exp^{\lambda BG_t}}$$

The estimated parameters for this variant are reported in the following table.

Table 7 Parameters for variant2 of similarity-based reinforcement learning model

ρ_U	ρ_F	δ_1	δ_2	BR_{ini}	λ	L
0.39	0.54	0.00	0.04	0.46	4.99	7610.6
[0.34, 0.44]	[0.49,0.58]	[0, 0.0007]	[0.03, 0.05]	[0.39, 0.52]	[4.8, 5.17]	-

In this variant, we see a slight discount for Red_2 but not for Red_1. The likelihood for this model is better than for the original model and the hypothesis $\delta_1 = \delta_2 = 0$ is rejected under significance level 0.01. While this extension has a slightly better explanatory power, we find only a modest level of ambiguity aversion applied to the Red urn in state 2 when allowed to differ from the ambiguity aversion to the Red urn in state 1.[24]

Generalized Bayesian learning model

For the Bayesian model, one could argue that instead of fixing $v(Blue) = 0.3$ and $v(Green) = 0.7$, the values of the Blue and Green urns could be updated similarly to the Red urns.[25] We have estimated such an extended model taking the same prior parameterized by the support $[k_{\inf}, k_{\sup}]$ for all the urns (see Table 8).

This model performs better than the generalized Bayesian one in terms of likelihood. However, this extended model is still statistically dominated by the similarity-based reinforcement learning model.[26]

A more elaborate version of the Bayesian approach would be to take into account the probability of having r_i state i in the 20 observation of the feedback (instead of assuming that in each round, there are exactly 10 subjects assigned to each state). Accordingly, we now represent the feedback as (b, g, n) where b, g are the number of draws from blue and green urns and n is the number of black balls in *Red*. We modify the generalized Bayesian

[24]We also considered the possibility that subjects would use a different slope to appreciate payoffs above 0.5 and payoffs below 0.5 in the spirit of prospect theory (with a reference payoff fixed at 0.5), but such a variant did not result in an improvement of the likelihood, hence we do not report it here (see Tversky and Kahneman (1974), (1979) for an exposition of prospect theory).

[25]The initial information provided about those urns would of course be used

[26]The Vuong test was conducted with the null hypothesis that both models explain the data equally well. The null was rejected in favor of the similarity-based reinforcement learning model with V_{stat}= 24.72.

Table 8 Parameter for Bayesian model with Blue and Green updating

λ	k_{inf}	k_{sup}	δ	Likelihood
8.69	3	7	0.008	8583.8
[8.4, 8.9]	(-)	(-)	[0.003, 0.013]	(-)

model by taking into account the probability of having x draws corresponding to states $s = 1$ out of 20 draws. Formally,

$$Pr(b, g, n|k_1, k_2) = \sum_x \left(\begin{array}{c} 20 \\ x \end{array} \right) \left(\frac{1}{2^{20}} \right) Pr(m_1 = x - b, m_2 = 20 - x - g, n|k_1, k_2)$$

where x is the number of times state $s = 1$ was observed in one round, $Pr(m_1 = x - b, m_2 = 20 - x - g, n|k_1, k_2)$ is defined as in section 4.2.2 where the total number of players in each session is 20.

The dynamics of beliefs is now given by

$$\beta_{t+1}(k_1, k_2) = \frac{\beta_t(k_1, k_2) \cdot \Pr(b, g, n|k_1, k_2)}{\sum_{r_1, r_2} \beta_t(r_1, r_2) \cdot \Pr(b, g, n|r_1, r_2)}.$$

with $\beta_1(k_1, k_2) = \beta_{init}(k_1, k_2)$. The other ingredients of the Bayesian learning model are identical to those considered in section 4.2.2.

After running the estimation of this model (see Table 9), we note that the corresponding likelihood is further improved as compared to those obtained with the other two Bayesian models. However, even with the improved likelihood, the model still underperforms relative to the reinforcement model, and the Vuong test still statistically favors the similarity-based reinforcement learning model.[27]

Table 9 Parameter for elaborate Bayesian model

λ	k_{inf}	k_{sup}	δ	Likelihood
6.56	3	7	0	8584.4
[6.57, 6.64]	(-)	(-)	[0, 0.008]	(-)

The Bayesian learning model makes an assumption that the initial prior belief regarding the composition of Red_1 and Red_2 follows a uniform distribution. As a robustness check consider a family of triangular distributions centered around 5, which together with the uniform prior considered in the main analysis allows us to span a wider range of priors.

[27]The Vuong test was conducted with the null hypothesis that both models explain the data equally well. The null was rejected in favor of the similarity-based reinforcement learning model with V_{stat}= 20.18.

Reminding that k_i denotes the number of black balls in Red_i, the trinagular distributions specify priors such that for some ll and uu with $ll < 5 < uu$ and for any $(k_1, k_2) \in [ll, uu]^2$

$$\beta_{init}(k_1, k_2) = \begin{cases} \frac{4(k1-ll)(k2-ll)}{((uu-ll)(5-ll))^2} & \text{if } ll < k1 < 5 \ \& \ ll < k2 < 5 \\ \frac{4(uu-k1)(uu-k2)}{((uu-ll)(5-ll))^2} & \text{if } 5 < k1 < uu \ \& \ 5 < k2 < uu \\ \frac{4(uu-k1)(ll-k2}{(uu-ll)^2(5-ll)(uu-5)} & \text{if } uu < k1 < 5 \ \& \ ll < k2 < 5 \\ \frac{4(k1-ll)(uu-k2)}{(uu-ll)^2(5-ll)(uu-5)} & \text{if } 5 < k1 < uu \ \& \ 5 < k2 < uu \end{cases}$$

and $\beta_{init}(k_1, k_2) = 0$ otherwise with the values of ll and uu to be estimated.

After running the estimation (see Table 10), we note that the uniform prior leads to a better fit so that our main finding that the generalized reinforcement learning model outperforms the Bayesian learning model would a fortiori be true had we considered the triangular priors.

Table 10 Parameter for triangle distribution Bayesian model

λ	ll	uu	δ	Likelihood
4.297	3	7	-0.019	9948.9
[4.29, 4.29]	(-)	(-)	[0, -0.0196]	(-)

6.2 Imitation heuristics

Our setting is about an individual decision environment in which every subject receives the same information/feedback. So if a subject trusts no less her/his ability to process the data/feedback as compared to others, there is no reason for this subject to reason about how others made their choices. As an alternative, one may consider subjects who would not trust their ability to process data/feedback and would instead try to behave like others maybe with the premise that others have a better sense of how to make good choices. This line of thought leads us to consider the following imitation heuristics.

In our experiment, everyone can infer from the proportions of Blue and Green urns that appear in the feedback, how frequently Red_1 and Red_2 were chosen in the pool, in the last round. If a subject thinks others had a good reason to make their choices, such a subject may try to adapt her/his behavior in the current round to take inspiration from the observed frequency of choices made in the last round. Based on this consideration, we propose that the probability of choosing Red in the two states would satisfy the following.

Choice Rule:

$$p_{1t} = \lambda + (1 - 2\lambda)\frac{(10 - NB_t)^\alpha}{NB_t^\alpha + (10 - NB_t)^\alpha}$$

$$p_{2t} = \lambda + (1 - 2\lambda)\frac{(10 - NG_t)^\alpha}{NG_t^\alpha + (10 - NG_t)^\alpha}$$

where $\lambda = 0$ and $\alpha = 1$ would correspond to the heuristic in which the behavior is probabilistic and matches exactly the frequencies observed in the last round, and a smaller λ and/or a larger α indicate a greater propensity to follow the majority choice. The learning model just considered is parameterized by λ, α together with p_{1init} and p_{2init} defining the initial probabilities of choices of Red_1 and Red_2 in the first round. Assuming our subjects follow this heuristic provides the following estimations given our observed data.

Table 11 Parameter for imitation heuristic model

α	λ	p_{1init}	p_{2init}	Likelihood
0.864	0.097	0.62	0.18	8651.7
(0.006)	(0.007)	(0.043)	(0.046)	(-)

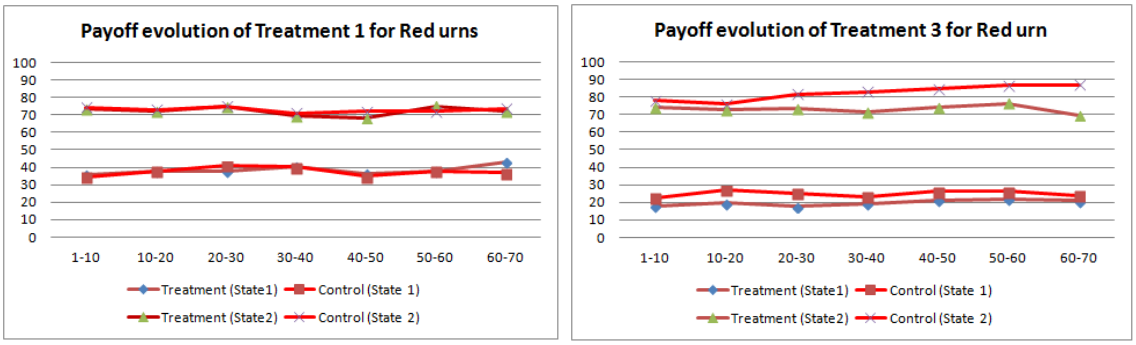
We observe that the estimation yields a heuristic model not very far from the standard matching model in which we would have $\lambda = 0$ and $\alpha = 1$, even if the fact that $\alpha = 0.864$ is indicative that in the estimated model behaviors are a bit less responsive than in the matching model. But, considering the likelihood, it appears that the imitation model provides a much less good fit than the similarity-based reinforcement learning model. We run a Vuong test comparing these two models at the aggregate level and found that the reinforcement model is a better fit to our experimental data than the imitation heuristic model (the Vuong test yields $V_{stat} = 6$).

In line with Section 4.4, we have also checked at the individual level, which of the imitation model or the Bayesian learning model or the reinforcement learning model allows offering the best prediction. That is, we compare at the individual level which of the reinforcement model with parameters in Table 3, or the generalized Bayesian model with parameters in Table 4 or the Imitation model with parameters in Table 11 explains best the observed choices of any given subject across the 70 rounds. We find that out of the 235 participants, 71.48% fit the Reinforcement model best, followed by 14.46% for the Imitation model and 14.04% for the Bayesian model. These statistical exercises reinforce the trust in the generalized reinforcement learning model as compared with alternative learning models in the context of our experiment.

6.3 Payoff comparison

Our study of the three main treatments is suggestive that some inefficiencies arise due to the coarseness of the feedback received by subjects. As we have also run control sessions for the return specifications of treatments 1 and 3 in which the feedback was made more detailed (separate feedback for Red_1 and Red_2), it is of interest to compare how the average payoffs compare between control and treatment over rounds. This is reported in the following figure.

As seems apparent the payoff is higher in control than treatment. To assess this from a statistical perspective, we have conducted a t-test. That is, we have compared the overall payoff as measured by the sum of black balls obtained over the 70 rounds across the population of subjects categorized as control or treatment. For the return specifications of Treatment 1 in which v_{Red_1}= 0.4 , v_{Red_2}= 0.8, the average sum was 38.2 black balls in control whereas it was 38.7 in treatment (even if small this difference is shown to be statistically significant with p-value 0.000). For the return specifications of Treatment 3 in which v_{Red_1}= 0.1 , v_{Red_2}= 0.9, the average sum was 37.68 black balls in control whereas it was 32.48 in treatment. This difference is bigger and statistically significant (p-value 0.000). Interestingly, the difference is much stronger for treatment 3 in which the large return of Red_2 is missed by several subjects. From this, we conclude that there can be substantial gains in providing detailed feedback if possible in such contexts.[28]



(a) Treatment1 **(b)** Treatment 3

Figure 6 Payoff difference between Control and Treatment

[28]The gains seem all the more pronounced that the returns to the unfamiliar choices are very heterogeneous and there is a strong negative correlation in the returns between the familiar and the unfamiliar choices, as is the case in treatment 3. More work is needed to confirm this more generally.

7 Conclusion

In this paper, we have considered the choices to be made between familiar alternatives and unfamiliar alternatives for which the obtained feedback is aggregated over different states of the economy. The literature on ambiguity aversion would suggest that the unfamiliar alternatives would be discounted as compared to the familiar ones, but that literature has largely ignored how behaviors would change in the face of continuously coming new feedback that would remain aggregated over different states.

Several competing learning models could be considered to tackle the choices in the face of new feedback: either extension of reinforcement models in the spirit of the valuation equilibrium (Jehiel and Samet, 2007) or Bayesian learning models in which subjects would start with some diffuse priors and update as well as they can, based on the coarse feedback they receive. Ideas of ambiguity aversion can be combined with such learning models along with the idea that subjects make noisy best-responses given their expectations, as routinely done in the empirical literature (discrete choice models as considered by McFadden) or in the experimental literature (quantal response equilibrium as defined by McKelvey and Palfrey, 1995).

Our results indicate that the similarity-based reinforcement learning models outperform their Bayesian counterparts and that little discount seems to be applied to unfamiliar choices even when the feedback relative to them is aggregated over different states, and as in other experimental findings, our results also indicate that subjects' choices require some noise in the best-response formulation. We believe such a work could be viewed as a starting point for a broader research agenda that aims at understanding how subjects make choices in the face of a mix of coarse and precise (state-specific) feedback. Understanding further whether and when subjects seek to generate state-specific feedback should also be part of this broader agenda.

References

[1] Barron, Kai Huck, S. and Jehiel, P. (2019). Everyday econometricians: Selection neglect and overoptimism when learning from others. *Working paper*.

[2] Camerer, C. and Ho, H.-T. (1999). Experience-Weighted Attraction Learning in Normal form games. *Econometrica*, 67.

[3] Cason, T. N., Sheremeta, R. M., and Zhang, J. (2012). Communication and efficiency in competitive coordination games. *Games and Economic Behavior*, 76:26–43.

[4] Charness, G. and Levin, D. (2005). When optimal choices feel wrong:a laboratory study of bayesian updating, complexity, and affect. *The American Economic Review*, 95(4):1300–1309.

[5] Cheung, Y.-W. and Friedman, D. (1997). Individual learning in normal form games:some laboratory results. *Games and Economic Behavior*, 19:46–76.

[6] Cownden, D., Eriksson, K., and Strimling, P. (2018). The implications of learning across perceptually and strategically distinct situations. *Synthese*, 195:511–528.

[7] Ellsberg, D. (1961). Risk, ambiguity and the savage axioms. *The Quarterly Journal of Economics*, pages 643–661.

[8] Enke, B. (2019). What you see is all there is. *Working paper; Harvard University.*

[9] Enke, B. and Zimmermann, F. (2019). Correlation neglect in belief formation. *The Review of Economic Studies*, 86:313–332.

[10] Epstein, L. and Halevy, Y. (2019). Hard-to-interpret signals. *Working Paper 634 University of Toronto.*

[11] Epstein, L. and Schneider, M. (2007). Learning under ambiguity. *Review of Economic Studies*, 74:1275–1303.

[12] Erev, I. and Roth, A. E. (1998). Predicting how people play games : Reinforcement learning in experimental games with unique mixed strategy. *The American Economic Review*, 88(4):848–881.

[13] Esponda, I. (2008). Behavioral equilibrium in economies with adverse selection. *American Economic Review*, 98(4):1269–91.

[14] Esponda, I. and Vespa, E. (2019). Contingent preferences and the sure-thing principle: Revisiting classic anomalies in the laboratory. *Working paper.*

[15] Fryer, R. and Jackson, M. O. (2008). A categorical model of cognition and biased decision making. *The B.E. Journal of Theoretical Economics*, 8(1):1–44.

[16] Fryer, R. G., Harms, P., and Jackson, M. O. (2019). Updating beliefs when evidence is open to interpretation: implications for bias and polarization. *Journal of European Economic Association forthcoming.*

[17] Fudenberg, D. and Levine, D. K. (1998). *The theory of learning in games.* MIT Press.

[18] Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18:141–153.

[19] Gilboa, I. and Schmeidler, D. (1995). Case based decision theory. *The Quarterly Journal of Economics*, 110:605–639.

[20] Grimm, V. and Mengel, F. (2012). An experiment on learning in a multiple games environment. *Journal of Economic Theory*, 147(6):2220–2259.

[21] Hopkin, E. (2002). Two competing models of how people learn in games. *Econometrica*, 70:2121–2166.

[22] Jehiel, P. (2005). Analogy-based expectation equilibrium. *Journal of Economic Theory*, 123(2):81–104.

[23] Jehiel, P. (2018). Investment strategy and selection bias: An equilibrium perspective on overoptimism. *American Economic Review*, 108(6):1582–97.

[24] Jehiel, P. and Koessler, F. (2008). Revisiting games of incomplete information with analogy based equilibrium. *Games and Economic Behavior*, 62(2):533–557.

[25] Jehiel, P. and Samet, D. (2005). Learning to play extensive games by valuation. *Journal of Economics Theory*, 121(557):129–148.

[26] Jehiel, P. and Samet, D. (2007). Valuation equilibrium. *Theoretical Economics*, 2(2):163–185.

[27] Ketz, P. (2018). Subvector inference when the true parameter vector may be near or at the boundary. *Journal of Econometrics*, 207(2):285–306.

[28] Martinez-Marquina, Alejandro Niederle, M. and Emanuel, V. (2019). Probabilistic states versus multiple certainties: The obstacle of uncertainty in contingent reasoning. *forthcoming American Economic Review*.

[29] McKelvey, R. D. and Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10:6–38.

[30] Mengel, F. (2002). Learning across games. *Games and Economic Behavior*, 74:601–619.

[31] Merkle, E. C., You, D., and Preacher, K. J. (2016). Testing non nested structural equation models. *Psychological Methods*, 21.

[32] Ryan, B. and Gross, N. (1943). Acceptance and diffusion of hybrid corn seed in two iowa communities. *Rural Sociology*, 8:15–24.

[33] Salmon, T. C. (2001). An evaluation of econometric models of adaptive learning. *Econometrica*, 69:1597–1628.

[34] Samuelson, L. (2001). Analogies, adaptations and anomalies. *Journal of Economic Theory*, 97:320–367.

[35] Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge Press.

[36] Train, K. (2003). *Discrete Choice Methods and Simulations*. Cambridge University Press.

[37] Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, New Series*, 185(4157):1124–1131.

[38] Tversky, A. and Kahneman, D. (1979). Prospect theory, an analysis of decision under risk. *Econometrica*, 47(2):263–291.

[39] Vespa, E. and Wilson, A. J. (2016). Communication with multiple senders: An experiment. *Quantitative Economics*, 7:1–36.

[40] Wilcox, N. (2006). Theories of learning in games and heterogeneity bias. *Econometrica*, 74:1271–1292.

[41] Young, H. P. (2009). Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *American Economic Review*, 99(5):1899–1924.

[42] Zimmermann, F. (2019). The Dynamics of Motivated Beliefs. *CRC TR 224 Discussion Paper Series; University of Bonn and University of Mannheim, Germany.*

8 Appendix

Appendix A: Instructions

Instruction sheet for the players (In the lab the instructions were in French):

Control Group:

Welcome to the experiment and I thank you for your participation. Please listen to these instructions carefully. If you have any questions kindly raise your hand and it shall be addressed. You receive 5 euros for participating and then your payoff depends on your performance in the experiment.

The Experiment:

The experiment consists of 70 rounds. It is a simple decision task. There are two situations you may face referred to as states 1 and 2. In each state, you have to choose one of two urns. Each urn is composed of ten balls either black or white in color. When you choose an urn, one of the balls in the urn is drawn at random (by the computer) and it is immediately replaced after the computer has noted the color of the ball. If the ball drawn is Black, you can receive extra payment (see below for details) whereas if the ball drawn is White you receive no payment.

The two urns available in state 1 are Blue and Red, respectively. The two urns available in state 2 are Green and Red, respectively. While the compositions of the various urns remain the same throughout the experiment, note that the compositions of the Red urn in

state 1 need not be the same as the composition of the Red urn in state 2. These are two different urns.

As the experiment goes, on your computer screen, you will be informed whether you have to make a choice of urns in state 1 (Blue or Red) or in state 2 (Green or Red). The sequence of choices from states 1 or 2 is decided randomly by the computer. Your task is to choose one urn out of the two in each state.

Note: We drew 100 times a ball (replacing the ball in the urn after each draw) out of the Blue and Green urn. We obtained the following composition

Blue	30 Black	70 White
Green	68 Black	32 White

At the beginning of the experiment:

- Your terminal is randomly assigned a State of the world. If in State 1, you choose between a Red and Blue urn. If in State 2, you choose between a Red and Green urn.
- After you choose the color of the urn that you want to pick, you click on the screen. A ball (the color of which could be either Black or White) will be drawn from that urn by the computer. You will not know the color of the ball drawn. This implies you will not have the information for your choice.
- Once all participants have made their choices, we provide you with some feedback. The total number of black and white balls drawn in previous rounds by all subjects according to the color of the urn (Blue, Red_1, Red_2, Green).
- Following the feedback, your terminal is randomly assigned a state of the world again. The state may vary from the previous round or remain the same.
- We then repeat the same experiment again until the completion of the 70 rounds.

For determining your payoff, two of the rounds will be randomly chosen at the end of the experiment. If one of your balls in these two rounds is Black, you will get an extra 5 euros. If both of your balls in these two rounds are Black, you will have an extra 10 euros. Otherwise (if both balls are White), you will have no extra return. So if you have no questions let us begin!

Treatment Group:

Welcome to the experiment and I thank you for your participation. Please listen to these instructions carefully. If you have any questions kindly raise your hand and it shall be addressed. You receive 5 euros for participating and then your payoff depends on your performance in the experiment.

The Experiment:

The experiment consists of 70 rounds. It is a simple decision task. There are two situations you may face referred to as states 1 and 2. In each state, you have to choose one of two urns. Each urn is composed of ten balls either black or white in color. When you choose an urn, one of the balls in the urn is drawn at random (by the computer) and it is immediately replaced after the computer has noted the color of the ball. If the ball drawn is Black, you can receive extra payment (see below for details) whereas if the ball drawn is White you receive no payment.

The two urns available in state 1 are Blue and Red, respectively. The two urns available in state 2 are Green and Red, respectively. While the compositions of the various urns remain the same throughout the experiment, note that the compositions of the Red urn in state 1 need not be the same as the composition of the Red urn in state 2. These are two different urns.

As the experiment goes, on your computer screen, you will be informed whether you have to make a choice of urns in state 1 (Blue or Red) or in state 2 (Green or Red). The sequence of choices from states 1 or 2 is decided randomly by the computer. Your task is to choose one urn out of the two in each state.

Note: We drew 100 balls randomly out of the Blue and Green urn which gave us the composition

Blue	30 Black	70 White
Green	68 Black	32 White

At the beginning of the experiment:

- Your terminal is randomly assigned a State of the world. If in State 1, you choose between a Red and Blue urn. If in State 2, you choose between a Red and Green urn.
- After you choose the color of the urn that you want to pick, you click on the screen. A ball (color of which could be either Black or White) will be picked up from that urn. You will not know the color of the ball drawn. This implies you will not have the information of your choice
- Once every participant has made their choice, we provide you with the feedback. The no. of black and white balls drawn from each colored urn (Blue, Red, Green) across states based on only the previous round draw is reported. *Note that the information for Red corresponds to the no. of balls picked in State 1 and 2.*
- Following the feedback, your terminal is randomly assigned a state of the world again. The state may vary from the previous round or remain the same. Note that the composition of the urn is however fixed throughout the experiment.
- We then repeat the same experiment again till we complete 70 rounds.

For determining your payoff, two of the rounds will be randomly chosen at the end of the experiment. If you have picked up B in that particular round, you end up with 5 euros more for each B otherwise no returns. So if you have no questions let us begin!

Appendix B: Monte Carlo Simulation for reinforcement learning model

The learning model we described is parameterized by $(\rho_R,\ \rho_B,\ \delta, \lambda,\ BR_{init})$. The diagrams below show that these parameters are normally distributed via Monte Carlo simulations with 1000 iterations and n=240. [29]

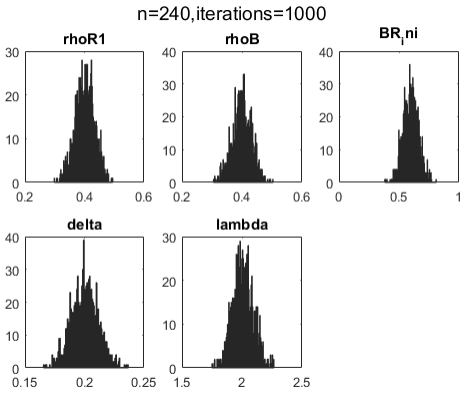
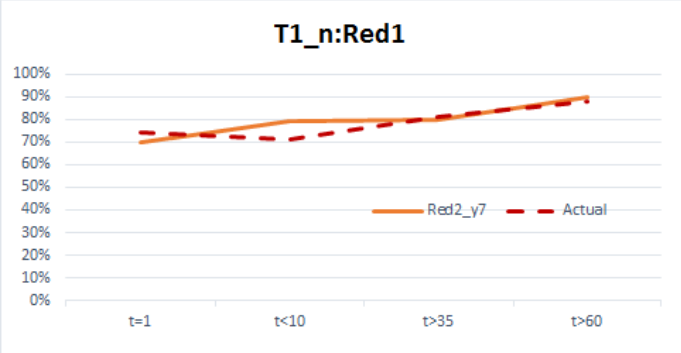


Figure 7 *Results for Montecarlo simulations for Model 1.*

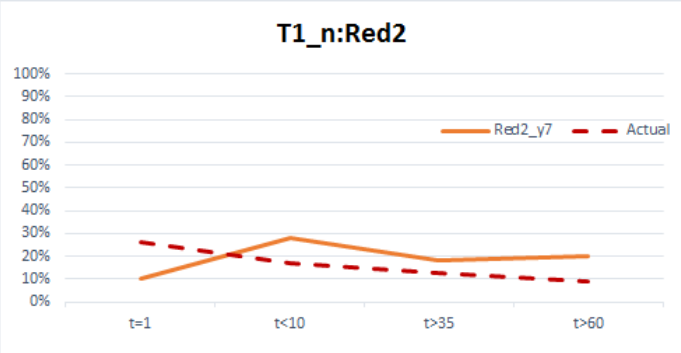
[29] This is in line with number of players in our actual experiment.

Appendix C: Simulated reinforcement with different noise parameter for T1

The figure shows simulated proportions of choices for the reinforcement model over 70 rounds with the estimated parameters for Treatment 1. Instead of using the noise parameter, $\lambda = 5.23$, we use $\lambda = 7$ to introduce less noise. This improves the fit of the simulated data with the actual one.



(a) Treatment 1: State 1

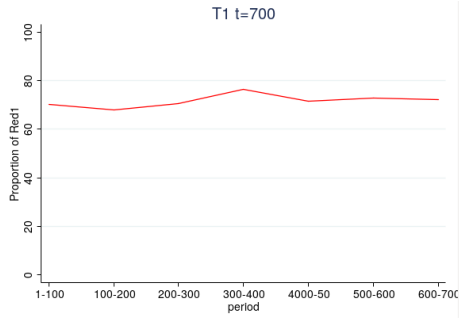


(b) Treatment 1: State 2

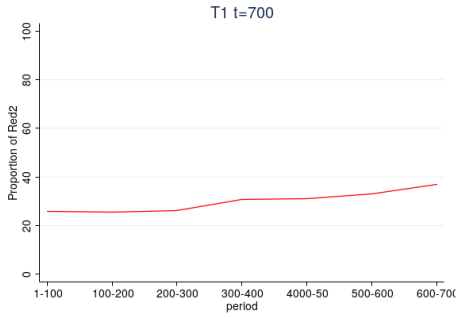
Appendix D: Drivers affecting Reinforcement learning

D.1 Incomplete learning

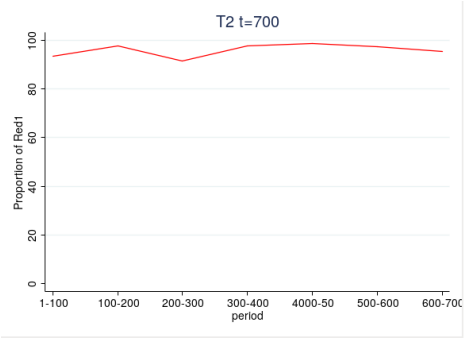
It is of interest to understand the gap between valuation equilibrium and optimal behavior is generated. The main factor driving this difference is the nature of the feedback being aggregate. One could argue that with 70 rounds, incomplete learning might be a concern. In Figure 12, we report the simulated frequencies of urn choices using the reinforcement model across all time periods and treatments using estimations from Table 2. Across all these sessions, our simulated frequencies show that the frequencies remain stable and there seems to be convergence at t=70. We can therefore rule out incomplete learning as a possible cause for the gap between valuation equilibrium and optimal behavior.



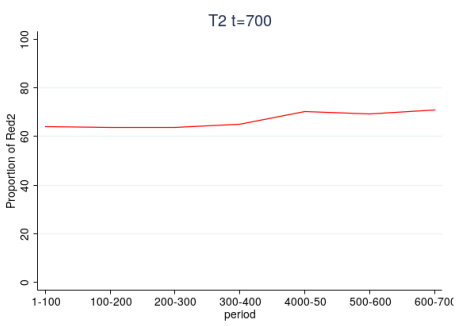
(a) Treatment 1: State 1



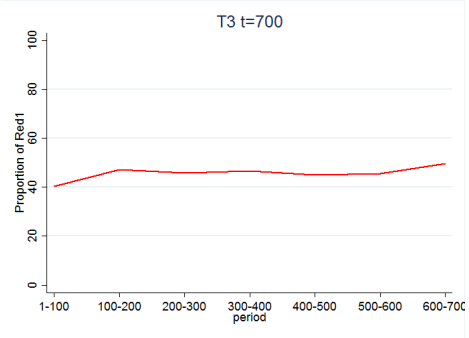
(b) Treatment 1: State 2



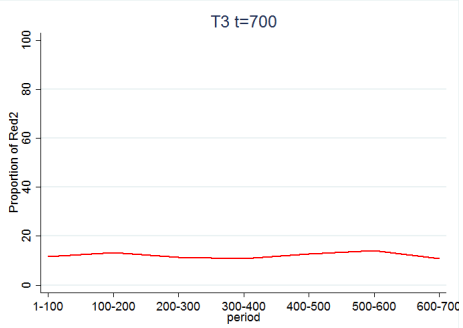
(c) Treatment 2: State 1



(d) Treatment 2: State 2



(e) Treatment 3: State 1



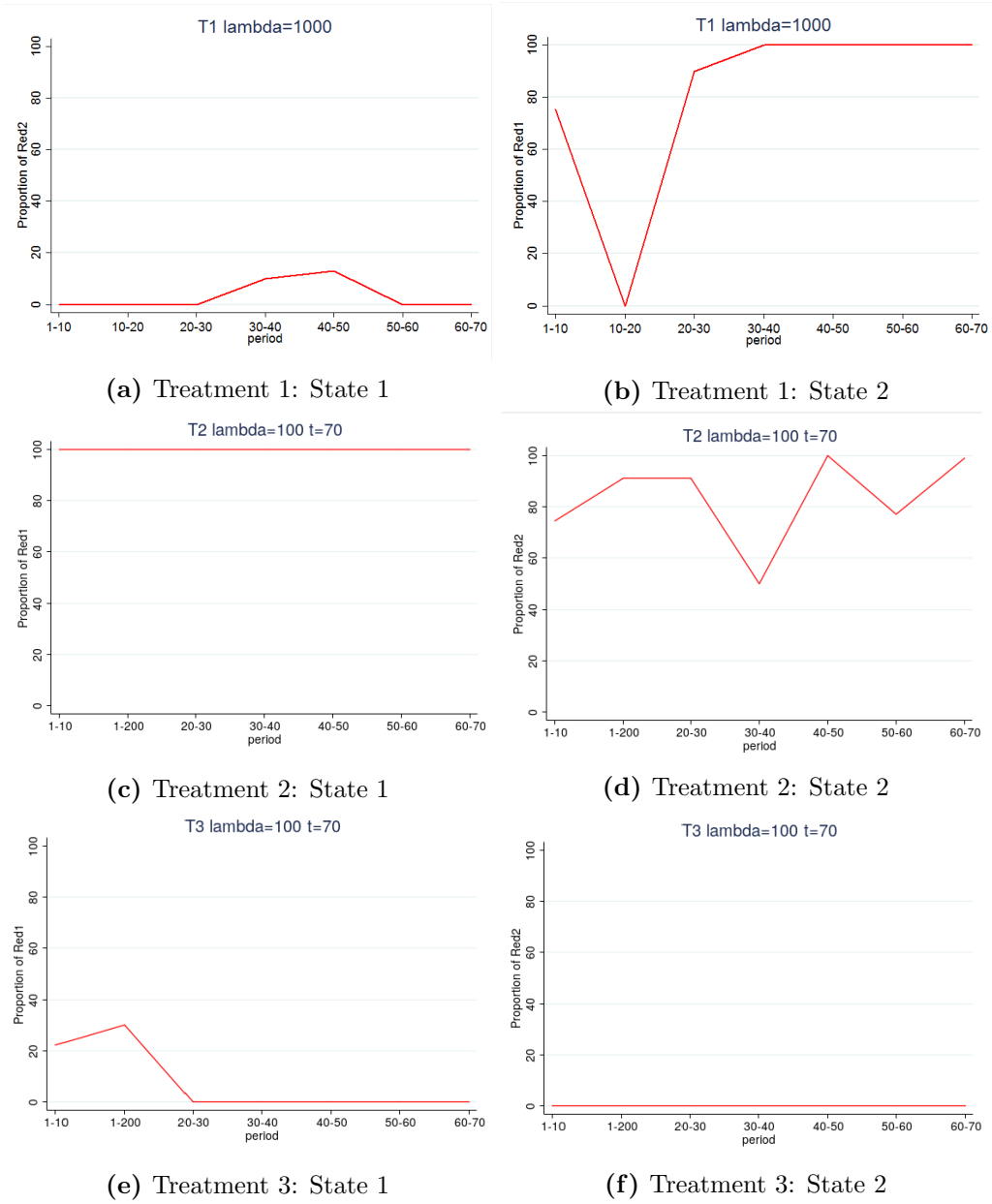
(f) Treatment 3: State 2

Figure 9 Simulated choices for t=700 using the reinforcement model

D.2 Noise

Another factor leading to the gap between optimal and valuation learning could be the level of noise represented by λ in Table 2. In Figure 13, we report the simulated frequencies of urn choices using the reinforcement model across all time periods and treatments using estimations from Table 2 and decreasing the level of noise by assigning λ=100. In both T2 and T3, the choices converge 100% to the ones predicted by Valuation equilibrium.

Consistent with Appendix C, a lower level of noise is required with α=1000 to achieve convergence to Valuation equilibrium.



(a) Treatment 1: State 1

(b) Treatment 1: State 2

(c) Treatment 2: State 1

(d) Treatment 2: State 2

(e) Treatment 3: State 1

(f) Treatment 3: State 2

Figure 10 Simulated choices for λ=100 (no noise) using the reinforcement model