Observational and Reinforcement Pattern-Learning: An Exploratory Study

Nobuyuki Hanaki
Alan Kirman
Paul Pezanis-Christou

SFI WORKING PAPER: 2017-07-022

SANTA FE INSTITUTE

Observational and reinforcement pattern-learning: An exploratory study[*]

Nobuyuki Hanaki[†] Alan Kirman[‡] Paul Pezanis-Christou[§]

July 17, 2017

Abstract

We examine, experimentally and theoretically in a very simple multi-armed bandit framework, how individuals learn about an undisclosed inter-temporal payoff structure. We propose a baseline reinforcement learning model that allows for pattern-recognitions and associated change in the strategy space, as well as its three augmented versions that accommodate observational learning from the actions and/or payoffs of another player with whom they are matched. The models reproduce the distributional properties of observed discovery times well. Our study further shows that observing another's actions and/or payoffs improves discovery compared to the baseline case when one of the pair discovered the hidden pattern.

Keywords: multi-armed bandit, reinforcement learning, payoff patterns, observational learning

JEL Code: D81,D83

[*] An earlier version of this paper was circulated under the title of "Counter intuitive learning: An exploratory study." We thank Ido Erev, Mamoru Kaneko, Rajiv Sethi, Christos Ioannou, Aidas Masiliunas, members of the Behavior Discussion Group at the Santa Fe Institute, and audiences at SAET 2014 (Tokyo), ASSET 2014 (Aix-en-Provence), and the Universities of Nice (GREDEG) and Toulouse for useful comments and suggestions, as well as Jade Wong and Justin Cheong for assistance with conducting the experiments. Support from the School of Economics of the University of New South Wales where the experiments were conducted, a LABEX OT-MED grant (ANR-11-LABX-0061), a JSPS-ANR bilateral "BECOA" grant (ANR-11-FRJA-0002), a ANR ORA-Plus "BEAM" grant (ANR-15-ORAR-0004) and the Australian Research Council (DP140102949) are gratefully acknowledged. The usual disclaimer applies.

[†] Université Côte d'Azur, CNRS and GREDEG. Email: nobuyuki.hanaki@unice.fr

[‡] CAMS, EHESS, and Aix Marseille University. Email: alan.kirman@ehess.fr [Corresponding author]

[§] School of Economics, University of Adelaide. Email: paul.pezanis-christou@adelaide.edu.au

1 Introduction

When individuals have to learn which of the many courses of action open to them yield the best results, they typically start exploring the various possibilities that they perceive. They start with a set of possible actions and try to choose amongst them but may learn, for example, that the payoff from an action taken today may depend on actions taken previously. They may also learn that there are other options available to them as they proceed.

Learning is necessary when the payoffs from taking actions are unknown. Most of the emphasis in the learning literature has been on reinforcing the probability of taking actions which have produced positive results (Bush and Mosteller, 1951) whether the reinforcement is based on the realized payoffs (Erev and Roth, 1998), the foregone payoffs (Camerer and Ho, 1999), or their regrets (Marchiori and Warglien, 2008). In this context, taking an action which produces a higher payoff than its average in the past should lead to taking that action with a higher probability in the future. However, there is an extensive literature which argues that individuals should also explore alternatives other than those which they have tried in the past even if some of the latter have given positive results. This is what is referred to as the "exploration-exploitation trade-off" (see, for example, Hills et al., 2015, for a review). The idea is that the individual weighs up the gain that he might get from trying different options against the value of pursuing previously successful actions. In this literature, it is generally assumed that the success obtained from taking an action is independent of the choices of other actions in the past.

Here we will consider the case in which the success of an endeavour depends on shifting from a currently successful action to another. In terms of the previous discussion this might seem counterintuitive but some simple examples may illustrate the point. The simplest case is that of fisheries. Having fished at one period in a certain area, fishermen move on to other areas knowing that otherwise their catch will diminish and that the remaining fish will not be sufficient to replace the current population. A more subtle problem, but still a classic example, is crop rotation where sustainable success depends on changing the crop cultivated each year. The yield from a crop grown on a given plot of land this year depends on what was grown there in the past. To benefit from this and to develop inter-temporal strategies to do so, the farmer has to become aware of the existence of these inter-temporal dependencies in the payoffs across various options. Once this happens he will then have to explore among the many possible inter-temporal choice strategies to discover the ones that result in higher payoffs. This problem is particularly difficult because farmers may well learn quite quickly that leaving the land fallow will improve the yield of the crop which was planted previously, but it is more complicated to understand that growing other crops on the same land may actually improve the yield compared to what it would have been if it had been left fallow. The process by which people become aware of such features of the environment that they have not known before, and learn to adopt potentially more complex, but better, behavioural strategies is not yet very well understood.

In this paper we use laboratory experiments to investigate whether and how individuals gain a better understanding of their environment and, as a result, obtain better outcomes. We are concerned with a situation in which individuals take actions and try to improve their performance but do not know whether there is necessarily a "solution", that is an optimal choice. What interests us, in particular, is how agents come to decide to try alternative actions to those which they have already used.

Two approaches can be adopted. One is to take the "impartial observer" who knows and understands the structure of the problem and then to find out how he would optimally find the solution to that problem if one exists. Then one would see if, in experiments, subjects' behaviour reflect that optimal behaviour. The second approach, which we will adopt here, is to observe what subjects do and then to develop a model which captures their learning behaviour without asking whether they achieve, or even try to achieve, an optimum.

We study a case in which participants are faced with a multi-armed bandit, the payoffs of which are correlated, and have an undisclosed temporal structure. Our experiment is thus different from other investigations of "standard" stationary multi-armed bandit problems in which all the arms generate stochastic payoffs from predetermined distributions, and the task for a subject is to find which one to choose (see, for example, Banks et al., 1997; Brown et al., 2009; Efferson et al., 2007; Hu et al., 2013; McElreath et al., 2005; Steyvers et al., 2009). In many cases in that framework, agents will optimally try arms to form an idea as to their expected payoffs from each arm and will have a stopping rule which tells them when to stop experimenting and stick to that arm which has proved most successful up to that point (see, e.g., Garivier and Kaufmann, 2016).[1]

As this approach does not suit a framework like ours (i.e., where payoffs have an undisclosed temporal structure), we develop and assess a reinforcement learning model that accounts for payoff patterns and that *(i)* allows a change in one's perception of the environment and *(ii)* accounts for the observation of another participant's actions and/or payoffs. We find that such a model organises the observed behaviour remarkably well. First, in the absence of information about others' choices or payoffs, this model clearly outperforms the standard reinforcement learning model in reproducing the observed distributions of discovery times and total payoffs. Second, variants of this model that incorporate information on another participant's last action or payoff suggest a reduced discovery time (or equivalently, a larger total payoff) that is qualitatively supported by the observations. However, the effect of this information is salient only for participants who discovered the pattern late (i.e., after the participant to whom they were matched). Third, the model's prediction that learning accelerates mostly when participants are provided information on both another's actions and payoffs is borne out by the data.

We briefly review the literature that is most relevant to our investigation in Section 2. Section 3 outlines the experimental design and procedures used. We present our baseline learning model that incorporates a basic pattern recognition feature and assumes no information about others' actions and/or payoffs in Section 4 and we report on its fit to the experimental data. In Section 5, we extend this baseline model to accommodate observational learning and check its out-of-sample performance in fitting the experimental data when observations are possible. Section 6 concludes.

2 Related literature

The type of situation we investigate relates to different streams of the literature on learning in complex environments. How individuals learn which action to take and how to condition that

[1] There is considerable work showing that the Gittens rule or others such as the Chernoff stopping rule are optimal for a large class of bandit problems. For a general discussion of the optimality of stopping rules even when there is dependence between the payoffs from successive use of the same arm, see Fryer and Harms (2017).

choice on previous experience has been a subject of considerable interest in the field of machine learning. There, a problem which corresponds to our simplest (baseline) case, is what is referred to as the "contextual bandit problem." As Agarwal et al. (2014) point out, contextual bandit problems are found in many important applications such as online recommendation and clinical trials. In this, as in our experiments, an agent collects rewards for actions taken over a sequence of rounds. At each point in time, the agent chooses an action to take on the basis of two things: the context for the current round, and the feedback, in the form of rewards, obtained in previous rounds. That literature has focused on the optimal choice of the probabilities with which to explore different actions (see, e.g., Auer, 2002; McMahan and Streeter, 2009; Beygelzimer et al., 2011).

Two important features of our investigation are the sequential nature of the payoffs and their undisclosed structure. The important change that is necessary for one to learn how to improve one's gain, is to recognise that there is a sequential pattern in the payoffs and then reinforce on sequences of actions and not on single actions. Thus one has to change the space of possible actions to one of actions conditioned on what has happened in the past, and this, in our framework, is the "context" in the machine learning literature to which we have just alluded.[2]

In our particular case, the recognition that sequences matter amounts to detecting patterns in the payoff sequence. This has recently been dealt with in a game theoretic context by Spiliopoulos (2012, 2013) who shows that introducing an ability to recognize patterns in the opponent's choices in belief-based learning model greatly improves the fit of the model to experimental data.[3] Recall, however, that the structure of payoffs in games is usually perfectly known to agents and conditions the search for patterns in the opponent's behaviour.[4] In our case, participants can be thought of

[2] The machine learning literature has also long dealt with analysing problems in "sequence extrapolation" (see, e.g Laird and Saul, 1994). That literature typically tries to develop deterministic algorithms that may be used for successful sequence predictions in general.

[3] See also the earlier work of Sonsino (1997) who motivates his idea of "learning to learn" with the following simple example. "Consider the case where three strategy profiles, say C, A, B, have been played repeatedly, for a long time, in that specific order, so that the history of play at some stage is . . . C, A, B, C, A, B, C, A, B, C, A, B, C, A. We claim that the players must "recognize" the repeated pattern if it has been repeated successively with no interruptions a large enough number of times."

[4] Such knowledge also makes it impossible to study if and how individuals end up learning the "correct model." There have been a number of accounts showing that individuals can have a wrong model in mind, but, as they learn what to do within that framework, their actions lead them to believe that the model is correct. Furthermore, their actions lead to the results that they expect and payoffs are, themselves, modified by the actions taken by the protagonists (see, e.g., Arrow and Green, 1973; Kirman, 1975, 1983; Bray, 1982; Woodford, 1990).

as playing “a game against nature” where nature is playing a deterministic sequence so that the problem boils down to one of identifying a structure in the payoffs without considering that they are generated by an “opponent.”

Another important stream of related research is the one on “observational learning” which came to the forefront with Bandura and McDonald (1963) and Bandura et al. (1963) who argued that the process of learning is greatly influenced by observing the behaviour of other individuals. Fryling et al. (2011) provide a general review of observational learning in the behavioural science literature whereas Smith and Sørensen (2011) focus theirs on observational learning in games with private information that entail herding.[5] As Bossan et al. (2015) show, and which is intuitive, imitation of good performance may do better than individual learning. This is simply because imitating successful individuals accelerates the learning process for the less successful.[6]

Our study relates most closely to that of Nedic et al. (2012) who faced subjects with a two-armed bandit whose arms, unknown to the subjects, paid off depending on the number of times that the arms had been chosen (the authors considered four payoff structures/treatments to investigate behaviour in varying environments). In these experiments, like ours, the payoff pattern of the bandit’s arms is unknown and subjects receive feedback on others’ choices, rewards or both.[7] They find that in some cases, observing the others’ payoffs (choices) may impede (accelerate) learning whereas observing both others’ payoffs and choices significantly accelerates it. An important difference with our setting is that, in some treatments, subjects did not need to discover the arms’ payoff structure to achieve the optimal payoff. In fact, once subjects realised that what was chosen previously had an impact, without knowing the precise form of that influence on their current payoff, they often tended to try to equalise the payoffs from the two arms and to switch when they achieved this. This is a rather natural heuristic but was optimal in only some of the treatments. As can be seen in our result section of our baseline treatment, we observed a similar phenomena in our experiments, such as, learned to use a heuristic which gave systematically good but not optimal

[5] See also Armantier (2004) who studies observational learning in a common value auction and shows that it speeds up convergence to Nash equilibrium bidding.

[6] They add, however, an important caveat: in a non-stationary environment, imitation may produce sub-optimal inertia in behaviour as imitation induces herd behaviour and discourages individual exploration.

[7] See also Burke et al. (2010) for a functional MRI study aimed at disentangling choice- from reward-based observational learning in an individual decision-making task.

results. We are particularly interested in studying how observational learning affects the discovery of an undisclosed payoff structure. We therefore consider a multi-armed bandit setting with the same information feedback configurations as Nedic et al. (2012). We now proceed to describe our experimental design.

3 Experimental design

One hundred and twenty-six students from the University of New South Wales participated in the experiment which consisted of four treatments, each consisting of 200 rounds of play followed by an incentivized questionnaire assessing the participants' risk preferences.[8] In each round, participants were individually asked to choose one of four options without being given any information about the possible payoffs they could generate. The underlying payoff generating process was such that the first three options generate a payoff of either 0 or 1 following a deterministic cycle whereas the fourth generates a constant payoff of 0.3. The payoff cycle of the first three options was such that in round t, option $a \in \{1, 2, 3\}$ generates a payoff of 1 if the remainder of $(t-1)/3$ equals $3-a$, otherwise the payoff is 0. That is, a payoff of 1 can be achieved in every round if the participant selects the right option a at the right time t, i.e., the optimal choice cycles in the order of 3, 2, 1, 3, 2, 1, 3,... from round 1, ..., 200. Recall again that the subjects did not know that the highest payoff was 1 nor that the payoffs followed a deterministic pattern.

Participants were randomly assigned to one of the four treatments which differed only in the amount of information feedback disclosed at the end of each round. In a baseline 'No Information' (NI) treatment, participants received no other information than the payoff outcome of their own choice. For the three other treatments, we used a partner matching protocol, i.e., participants were each randomly matched with one other participant and kept that partner for the 200 rounds of play, and we provided participants with some information about the choices and/or payoffs of the partner with whom they were matched. In a 'Choice Information' (CI) treatment, participants were informed about the other participant's last choice, but not the payoff outcome of that choice.

[8] The experiment was computerized and programmed in z-Tree (Fischbacher, 2007) The students were enrolled in Business Administration, Law, Marketing or International studies and were recruited by public advertisement on campus using ORSEE (Greiner, 2015).

Table 1: Summary of experimental sessions

Treatment	No. of subjects	Description
NI	30	Subjects play the game in isolation
CI	32 (16 pairs)	Subjects observe the last choice made by the other subject
PI	32 (16 pairs)	Subjects observe the last payoff outcome of the other subject
CPI	32 (16 pairs)	Subjects observe both the last choice and payoff of the other subject

Note: CI, PI and CPI used a ’partner’ matching protocol.

In a ‘Payoff Information’ (PI) treatment, they were informed about the other participant’s last payoff outcome, but not the corresponding choice. And in a ‘Choice & Payoff Information’ (CPI) treatment, they were informed about the other participant’s last choice and payoff outcome. Table 1 summarizes the experimental sessions.

Participants were not allowed to record any information in any form and had to rely on their memory of past play and outcomes to make their decisions. In addition, to reduce the inferences they could make from observing the other’s last choice and/or payoff outcome, they were not informed that they would receive the same payoff if they chose the same option as the other participant with whom they were matched. To incentivize decisions, individual payoffs (expressed in ECUs) were cumulated over the 200 rounds and converted into cash at the end of the experiment at the rate of 20 Cents per ECU.

Once the 200 rounds of play were over, and before cashing their rewards, all participants received a new set of instructions outlining the Holt and Laury (2002) lottery choice questionnaire which aims at measuring their risk preferences. Again, to incentivize decisions, one of the lottery choice tasks was randomly chosen once the questionnaire was completed and individual rewards for this second part of the experiment were determined by the lottery chosen in that particular task. The lottery payoffs were expressed in Australian Dollars.[9] We chose to elicit the participants’ risk averse preferences *after* they completed the experiment to avoid a possible framing effect that would lead them to perceive the payoffs in the bandit task as being random. The realised average individual

[9] A copy of the instructions is provided in the Appendix A.

reward from participating to the experiment, including the lottery choice questionnaire, was A$30 (which included a A$5 show-up fee) for a maximum of one hour and a half spent in the laboratory, including the time needed to read the instructions.

4 Learning alone

4.1 Descriptive data analysis

We start with a brief model-free overview of the most salient types of behaviour. Clearly, as participants did not know the exact structure underlying the payoffs the four options generate, many choice patterns could be believed to be 'optimal'. Here we focus on the participants' times of discovery of the hidden cycle generating a payoff of 1 in every round. We diagnose a discovery when a participant earns a payoff of 1 for twelve consecutive rounds and we define the time of the discovery as being the median round number of the first batch of twelve consecutive rounds generating a payoff of 1.[10] Out of the thirty subjects who participated in this treatment, exactly half of them discovered the hidden pattern but they did so at very different times: five of them found it within 50 rounds, eight of them found it between 51st and 100th round and two of them between 101st and 200th round. The other half of subjects in this treatment either never found the pattern or settled for the safe option (option 4) which generates a constant payoff of 0.3.

Figure 1 shows the time-series of choices and payoff outcomes for six participants who displayed different types of behaviour. While some discovered the hidden cycle within 50 rounds (Subject 19), others 'explored' for a few rounds and settled for the safe option. Subject 10, for example, never got a payoff of 1 when choosing other options while Subject 1 decided to exploit the safe choice eventhough s/he sometimes got a payoff of 1 when choosing other options. Some participants also settled into more complex and less profitable choice patterns: Subject 20 settled into a nine-round cycle repeating the choice sequence (3, 4, 4, 4, 4, 1, 4, 4, 4), Subject 25 settled into the six-round cycle (1, 3, 2, 4, 4, 4) while Subject 30 settled into another, almost optimal, six-round cycle (3, 2,

[10] For example, if someone receives a payoff of 1 from round 40 to 200, then his/her discovery time is round 45. As we will see, some participants occasionally 'deviated' from the hidden pattern after having exploited it for a long time, but most of them quickly returned to it.

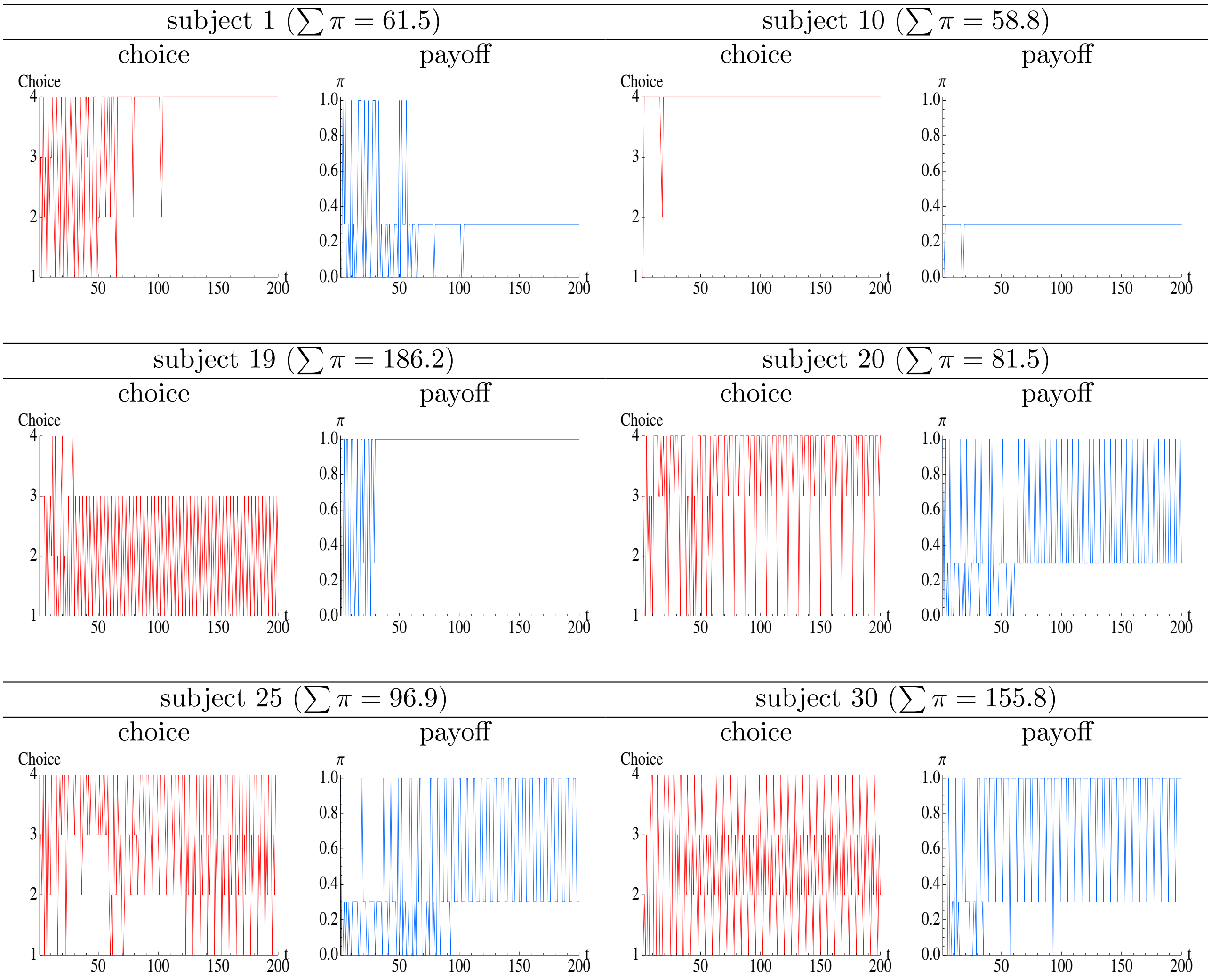
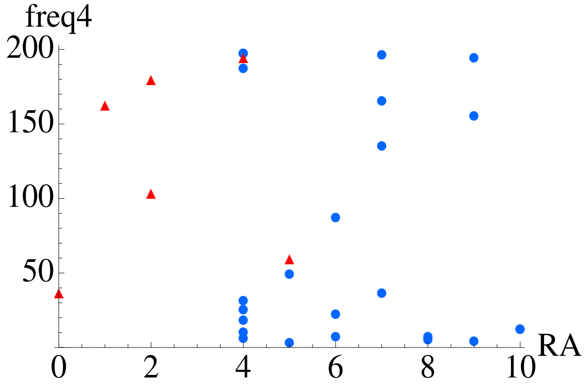


Figure 1: Examples of choice and payoff patterns in the baseline treatment.



Without triangle: N=24. $\rho = -0.148$ ($p = 0.496$)
Including triangles: N=30. $\rho = -0.260$ ($p = 0.167$)

Figure 2: Risk Aversion *vs* Frequency of Safe Option. Triangle: subjects who have switched multiple times in Holt and Laury (2002) questionnaire.

1, 3, 2, 4). These are heuristics that gave the sort of systematically good but not optimal results to which we have alluded in the end of Section 2 above.

One might reasonably conjecture that risk averse individuals are more likely to exploit the safe option than to explore risky ones, so we checked whether the individuals' frequencies of choosing the safe bandit is positively correlated with their aversion toward risk. Recall that the Holt and Laury (2002) questionnaire consists of a series of lottery choice tasks in which participants are asked to choose between two lotteries, one of which is less variable than the other. The lotteries' payoffs are such that when the probability of the High payoff for both options increases, an expected utility maximising agent should choose the less variable lottery in the early tasks and at some point switch to the more variable one; the later the task at which this switch occurs the more risk averse is the subject.[11]

Figure 2 shows a scatter plot of the degree of risk aversion (the subjects' switching points) versus the frequency of choosing the safe option (freq4) along with a Spearman correlation coefficient. As the reported coefficient is not statistically significant at $\alpha = 5\%$, we discard risk aversion as a

[11] Six out of thirty participants reverted to the "safe lottery" after having stopped choosing it. As such behaviour might be considered aberrant since it seems to be inconsistent with rational choice theory under risk, we conducted our analysis with and without these participants but found no significant change in our conclusions. For the analyses including these subjects, we have used their first switching point to measure their degree of risk aversion. It is also worth noting that in our context, if an individual who has chosen the safe arm, switches unsuccessfully to another choice it would not be irrational to come back to the safe arm.

possible explanation for the observed behaviour. This non-significant relationship between the observed behavior and risk preference is similar to that which Banks et al. (1997) report in their bandit experiments.[12] We now develop a reinforcement learning model that aims at reproducing the observed behaviour.

4.2 A pattern-learning model

We assume that agents, who are initially not aware of the possible payoffs that each option generates, remember the payoffs that result from choosing each option as soon as it is observed. Thus, for our simple problem, after enough trials, agents will learn about all the possible outcomes that each option generates. Let $\Pi_t^i(a)$ represent the set of payoffs agent i has observed each time he chose option a until round t. Thus, if agent i has experienced all the possible outcomes from choosing all four options at least once, before round t, we have $\Pi_t^i(a) = \{0, 1\}$ for $a \in \{1, 2, 3\}$ and $\Pi_t^i(4) = \{0.3\}$. Of course, agents cannot be certain that what they have observed are indeed all the possible outcomes.[13] In addition, we need to decide how the agents in our model make their choices amongst these different options.

The first step will be for individuals to recognise that their gain does not simply depend on the particular option chosen but also on what point in time that option is chosen as opposed to some alternative. Thus, the subjects in our experiments should not just condition their choices on past experience with those choices but also on the "context." In our case, the context is the order (or timing) in which options are chosen. To incorporate these considerations in our model, we need to make an assumption as to how agents become aware of the existence of a pattern in the payoff generating process. We assume that agents become aware of the possibility of the existence of such a pattern in the payoffs from options $a \in \{1, 2, 3\}$ if they observe the sequence (choice, payoff) of length $l^i \geq 2$, *starting with a payoff of 1*, $r^i \geq 1$ times. If $l^i = 2$ and $r^i = 2$, after observing the sequence of choice payoff pair $(1, 1), (1, 0)$, for example, over two consecutive rounds twice, agent i will start considering the existence of a dependency (namely, choosing the option 1 after choosing

[12] We also did not find a clear correlation between the measured degree of risk aversion and the observed frequencies of safe choices in the three paired treatments that are studied in the following section. See Appendix B.

[13] It is also possible that the time necessary to observe all possible outcomes could be long depending on how an agent explores. So the set of possible outcomes perceived by an agent could be less than the whole set.

it and obtaining payoff of 1 will result in payoff of 0).[14]

Once i starts considering such dependencies, he will start choosing the option conditional on the outcome in the previous round. In particular, i will start remembering the possible payoffs each conditional choice generates. Let $\Pi_t^i(a|h)$ represent the set of payoffs agent i has observed by choosing option a conditional on history h until round t. For example, $\Pi_t^i(1|(2,1))$ will be either empty or 1. While it is possible that an agent may condition his choice of options on the outcomes of two or more previous rounds, we restrict our attention to those choices which are conditioned only on the outcome of the most recent round (i.e., the choice of round t, a_t, depends on the outcome of round $t-1$ ($h_{t-1}^i = (a_{t-1}^i, \pi_{t-1}^i)$)). Because of this assumption, we also assume $l^i = 2$ for all i. Note that for the problem we consider in this paper, the set of all the possible outcomes in the previous round, h_{t-1}^i, is $\{(1,0),(1,1),(2,0),(2,1),(3,0),(3,1),(4,0.3)\}$.

How will these outcomes contribute to determining choices? We consider two sets of strategies: unconditional and conditional. Unconditional strategies, those used before a subject becomes aware of sequences, are simply choices of options $a \in \{1,2,3,4\}$. Conditional strategies are round t choices of options conditional on the outcomes in round $t-1$, $s = a|h$. Agents start using conditional strategies only after they have become aware of the existence of a sequential pattern as we have described above.

Let us first describe an unconditional strategy. $A_t^i(a)$ summarizes, at the beginning of round t, the past experience for agent i from choosing option a. Let $A_0^i(a) = 0.5$ for all i and a. We assume that $A_t^i(a)$ evolves as follow:

$$A_{t+1}^i(a) = \begin{cases} \alpha^i A_t^i(a) + (1-\alpha^i)\pi_t^i & \text{if } a = a_t^i \\ A_t^i(a) & \text{otherwise} \end{cases}$$

where a_t^i and π_t^i denote the option chosen by agent i in round t and the resultant payoff, and $\alpha^i \in (0,1)$ captures the weight put on past experience.

[14] We can easily allow an agent to start considering dependencies based on (choice, payoff) sequences that do not start with payoff of 1. This will in fact speed up the learning of the inter-temporal dependencies. We did not do so here because (choice, payoff) sequence starting with payoff of zero, such as $(1,0)$ can result in payoff of both 0 and 1 for all the possible choices among three options other than option 4.

Given $A_t^i(a)$, we assume that the probability of agent i choosing option a in round t is

$$Pr(a_t^i = a) = \frac{e^{\lambda^i A_t^i(a)}}{\sum_k e^{\lambda^i A_t^i(k)}}$$

where λ^i reflects the sensitivity of i's choice to his previous experiences. If $\lambda^i \to \infty$ he chooses the action which has the highest $A_t^i(a)$ and if $\lambda^i = 0$ he simply tries all actions with equal probabilities.[15]

Now let us describe i's choice among conditional strategies. Let $B_t^i(a|h)$ summarize i's experience from choosing a conditional strategy $a|h$ at the beginning of round t. Let $B_\tau^i(a|h) = A_\tau^i(a)$ for all h at round τ when i became aware of the inter-temporal dependencies. After that, $B_t^i(a|h)$ evolves as follows

$$B_{t+1}^i(a|h) = \begin{cases} \beta^i B_t^i(a|h) + (1-\beta^i)\pi_t^i & \text{if } a|h = a_t^i|h_{t-1}^i \\ B_t^i(a|h) & \text{otherwise} \end{cases}$$

where $\beta^i \in (0,1)$ captures the weight put on past experience.

Based on $B_t^i(a|h)$, the agent chooses option a in round t according to

$$Pr(a_t^i = a|h_{t-1}^i) = \frac{e^{\mu^i B_t^i(a|h_{t-1}^i)}}{\sum_k e^{\mu^i B_t^i(k|h_{t-1}^i)}}$$

where μ^i reflects the sensitivity of i's choice of a conditional strategy to past experiences. For simplicity, we assume $\alpha^i = \beta^i$ and $\lambda^i = \mu^i$ for all i.

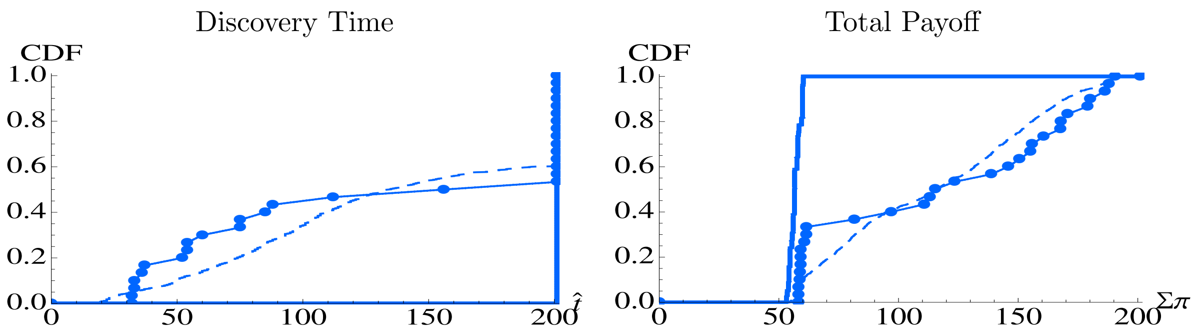
We fit our model to the experimental data for each subject by searching for a set of $(\alpha^i, \lambda^i, r^i)$ from the predefined parameter space that maximizes

$$\sum_{t=1}^{200} \ln(P_t^i(a_t^i))$$

where $P_t^i(a_t^i)$ is the probability of observing choice a_t^i according to the model described above.

The parameter space we consider is $\alpha^i \in [0.01, 0.99]$ with a step of 0.01, $\lambda^i \in [0.1, 20.0]$ with a

[15] This logistic rule is commonly used in learning models, and can be derived as an optimal trade-off between exploitation and exploration (Nadal et al., 1998; Bouchaud, 2013).



Exp: 133.23 (73.0), Sim: 135.88 (61.31)
p=0.814

Exp: 119.06 (50.20), Sim: 115.06 (40.74)
p= 0.597

Figure 3: CDFs of discovery times and total payoffs. Legend: Observed (thin solid with makers), Simulated baseline model (thin dashed), and Simulated constrained model (thick solid). Mean (standard deviation) for observed and simulated baseline model are also shown. P-values are based on permutation tests (for independent samples), two-tailed.

step of 0.1, and integer $r^i \in [1, 10]$. When the sum of log-likelihood has no unique maximum, we select the set of parameter with the smallest r^i as this parameter appears to be the unique source of non-uniqueness. Hence, we start by fitting the model to determine the set of best parameter values $\left\{(\hat{\alpha}^i, \hat{\lambda}^i, \hat{r}^i)\right\}$ for each individual i.[16] We then consider a population of 1000 artificial agents with parameters randomly drawn (with replacement) from this set and we compare the simulated outcomes to those observed in terms of individuals' discovery times and total payoffs.

Figure 3 displays the observed and simulated cumulative distributions (solid and dashed thin lines, respectively) of discovery times and total payoffs. Running permutation tests (for independent samples) do not reject the null of the stochastic equivalence of simulated and observed data samples ($p = .814$ for discovery times and $p = .597$ for total payoffs, two-tailed tests) so that the model fits the distributional properties of discovery time and total payoffs remarkably well. To put the goodness-of-fit of this pattern-learning model into perspective, we compare it to the one of a constrained version that disactivates the agent's ability to recognize patterns and to consequently use conditional strategies. This is achieved by imposing $r^i > 200$ (the number of rounds of play) which amounts to assuming that the agent never becomes aware of the existence of a temporal payoff

[16] See Appendix C for sample time series of choices and payoffs generated by the fitted model.

structure and thus never uses conditional strategies. We fit this constrained version by searching for each individual i the set of best parameters $\left\{(\tilde{\alpha}^i, \tilde{\lambda}^i)\right\}$. The resulting plots of the simulated distributions in Figure 3 (solid thick lines) clearly indicate no discovery within 200 repetitions and consequently lower payoffs than those observed or simulated by our baseline model.

5 Learning from others

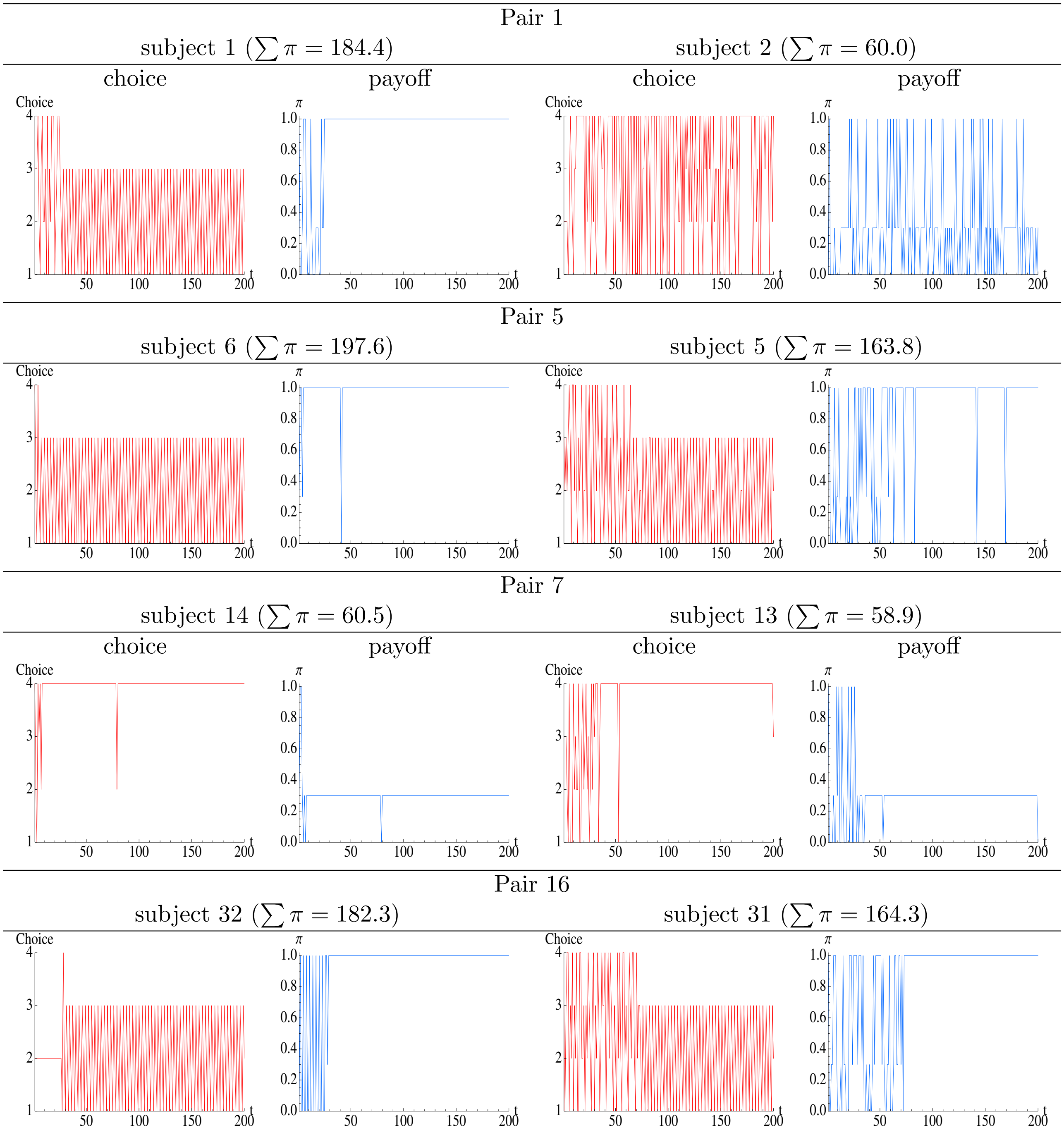
We now extend our baseline treatment and model to allow for and accommodate observational learning.

5.1 Descriptive data analysis

Figures 4, 5 and 6 display the time-series of choices and payoffs for four pairs of participants in each of the three additional information treatments considered, i.e., CPI, PI, and CI. The plots indicate that while some pairs never discovered the hidden pattern (Pair 7 in CPI and in CI and Pair 12 in PI), in a few others, one participant discovered it whilst the other did not (Pair 1 in CPI, Pair 3 and 14 in PI and Pair 5 in CI). Yet, in other pairs, both participants eventually discovered patterns (Pairs 5 and 16 in CPI, Pair 1 in PI and Pairs 1 and 4 in CI). As in the NI treatment, a few participants who discovered the pattern occasionally deviated from it in subsequent rounds (Subject 5 and 6 of Pair 5 in CPI, Subject 27 of Pair 14 in PI and Subject 19 of Pair 4 and Subject 23 of Pair 5 in CI). In contrast to the NI treatment, we find no evidence of participants settling into more complex and less profitable choice patterns.

5.2 Three observational pattern-learning models

We present three extensions of our baseline pattern-learning model that allow for observational learning. The three extensions assume that agents keep track not only of their own choices and payoffs over time but also part of, or all of the other's, and thus learn on the basis of these two information sets. However, as the assumption that agents pay equal attention to their own outcomes as to the other's outcomes has been challenged in the experimental literature (see, e.g., McElreath



Pair 1

subject 1 ($\sum \pi = 184.4$)

subject 2 ($\sum \pi = 60.0$)

choice payoff choice payoff

Pair 5

subject 6 ($\sum \pi = 197.6$)

subject 5 ($\sum \pi = 163.8$)

Pair 7

subject 14 ($\sum \pi = 60.5$)

subject 13 ($\sum \pi = 58.9$)

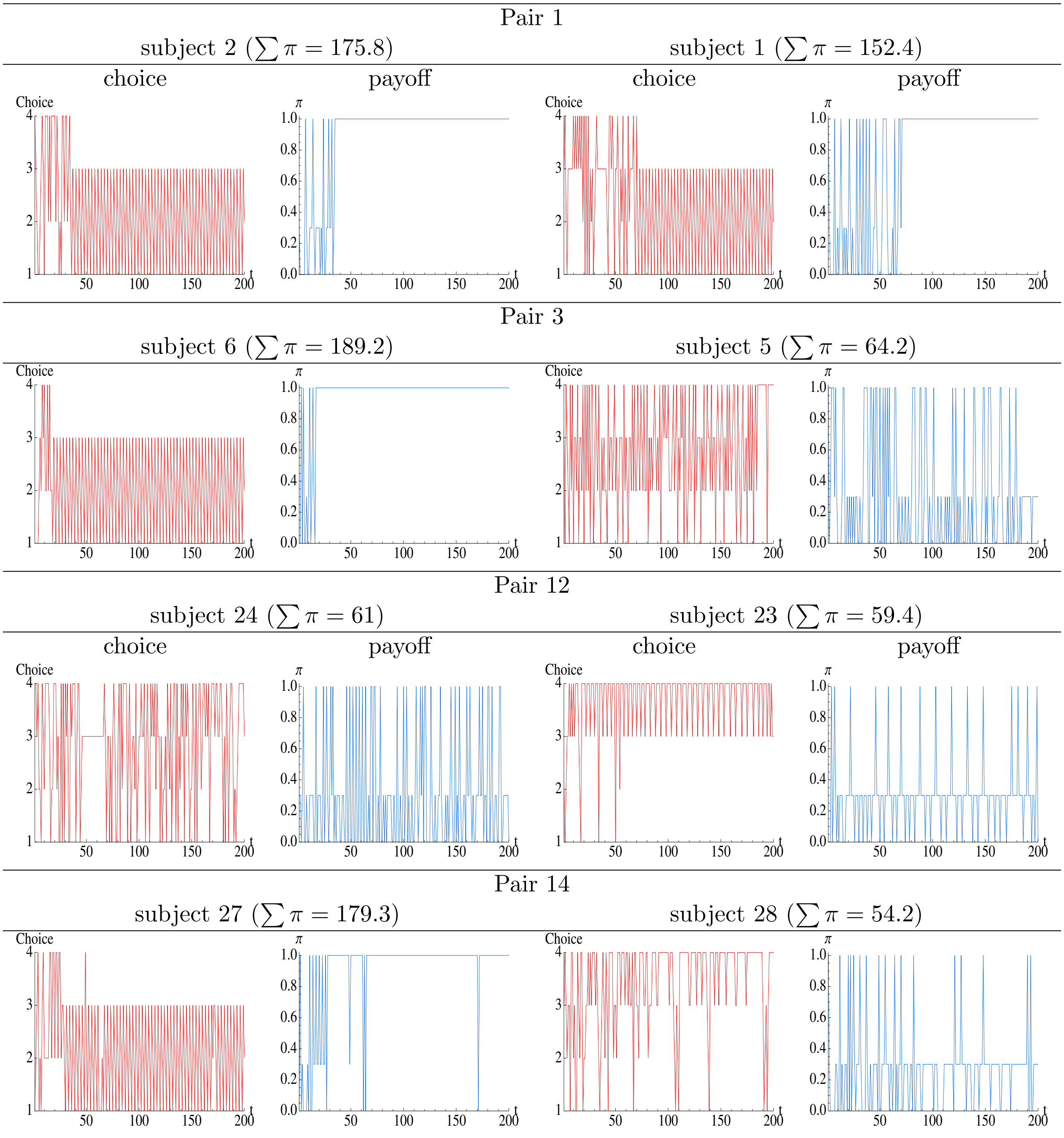
choice payoff choice payoff

Pair 16

subject 32 ($\sum \pi = 182.3$)

subject 31 ($\sum \pi = 164.3$)

Figure 4: Examples of choice and payoff patterns in the CPI treatment.



Pair 1

subject 2 ($\sum \pi = 175.8$) subject 1 ($\sum \pi = 152.4$)

choice payoff choice payoff

Pair 3

subject 6 ($\sum \pi = 189.2$) subject 5 ($\sum \pi = 64.2$)

Pair 12

subject 24 ($\sum \pi = 61$) subject 23 ($\sum \pi = 59.4$)

choice payoff choice payoff

Pair 14

subject 27 ($\sum \pi = 179.3$) subject 28 ($\sum \pi = 54.2$)

Figure 5: Examples of choice and payoff patterns in the PI treatment.

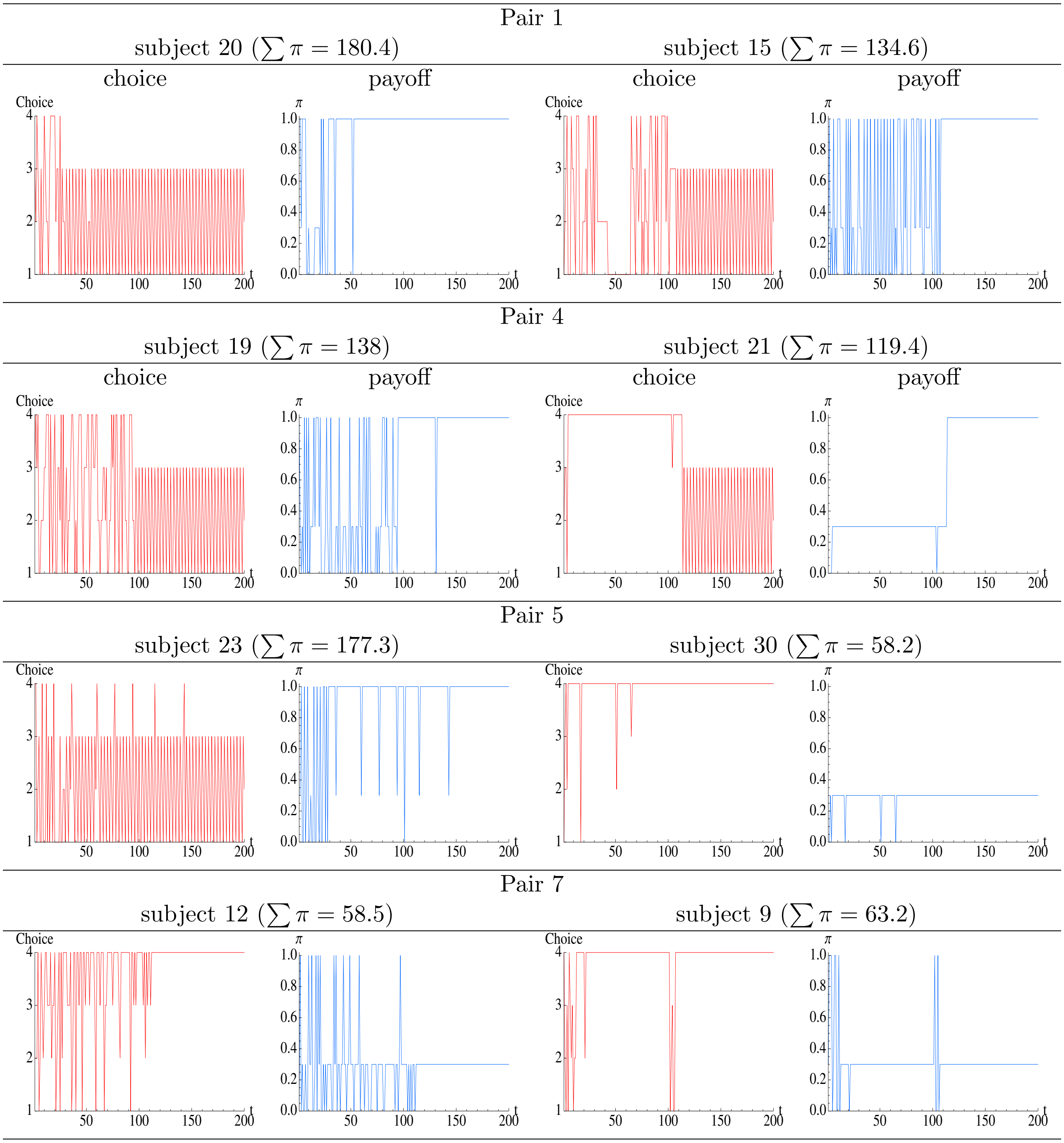


Figure 6: Examples of choice and payoff patterns in the CI treatment.

et al., 2005; Bayer and Wu, 2016), we assume that agent i pays attention to (a^i, π^i) for sure and pays attention to (a^j, π^j) with probability $q^i \in [0,1]$.[17] Let $\Gamma_t^i = 1$ represent i paying attention to (a_t^j, π_t^j) in round t (which happens with probability q^i) and $\Gamma_t^i = 0$ otherwise.

5.2.1 The 'Choice & Payoff Information' (CPI) case

Let $A_t^i(a)$ summarize, at the beginning of round t, the past experience for agent i from choosing option a. Let $A_0^i(a) = 0.5$ for all i and a. We assume that $A_t^i(a)$ evolves as follows:

$$A_{t+1}^i(a) = \begin{cases} \alpha^i A_t^i(a) + (1-\alpha^i)\pi_t^i & \text{if } a = a_t^i \\ \alpha^i A_t^i(a) + (1-\alpha^i)\pi_t^j & \text{if } a = a_t^j \neq a_t^i \text{ and } \Gamma_t^i = 1 \\ A_t^i(a) & \text{otherwise} \end{cases}$$

As for the NI treatment, we assume that agent i becomes aware of inter-temporal dependencies when i observes the same sequence of (choice, payoff) pairs. However, this sequence of (choice, payoff) pairs can now be based not only on the sequence of his own (choice, payoff) pairs but also on the observed sequence (assuming that $\Gamma_t^i = 1$ for two consecutive rounds frequently enough) of j's (choice, payoff) pairs.

Let $B_\tau^i(a|h) = A_\tau^i(a)$ for all h at round τ when i became aware of the inter-temporal dependencies. We assume, after round τ, that $B_t^i(a|h)$ evolves as follows:

$$B_{t+1}^i(a|h) = \begin{cases} \beta^i B_t^i(a|h) + (1-\beta^i)\pi_t^i & \text{if } a|h = a_t^i|h_{t-1}^i \\ \beta^i B_t^i(a|h) + (1-\beta^i)\Gamma_t^i\pi_t^j & \text{if } a|h = a_t^j|h_{t-1}^j \neq a_t^i|h_{t-1}^i \text{ and } \Gamma_t^i = \Gamma_{t-1}^i = 1 \\ B_t^i(a|h) & \text{otherwise} \end{cases}$$

[17] Note that keeping track of whether participants pay attention to the other's outcomes or not (for example by displaying the information proviso a mouse-click) might provide an estimate of q or allow some modelling of how it changes over time. However, since accessing this information does not guarantee that the participant factors it in her/his decision, it remains unclear what such an analysis would reveal.

Based on $B_t^i(a|h)$, the choice of option a for agent i in round t is thus be defined as:

$$Pr(a_t^i = a) = \frac{e^{\lambda^i B_t^i(a|h_{t-1}^i)} + \Gamma_{t-1}^i e^{\lambda^i B_t^i(a|h_{t-1}^j)}}{\sum_k e^{\lambda^i B_t^i(k|h_{t-1}^i)} + \Gamma_{t-1}^i \sum_k e^{\lambda^i B_t^i(k|h_{t-1}^j)}}$$

5.2.2 The 'Payoff Information' (PI) case

The basic idea here is that when i observes j obtaining a sequence of payoff 1s, i tries to figure out how j is obtaining such a stream of high payoffs based on (1) what i knows about the set of possible (choice, payoff) pairs and (2) their inter-temporal dependencies. On the one hand, if i is not yet aware of the existence of the inter-temporal dependencies and conditional strategies, we assume that by observing j's sequence of high payoffs, i becomes aware of the existence of such dependencies and strategies in addition to the process that we have described in our baseline model based on sequences of agent's own choice-payoff pairs. Adopting the same reasoning as in the case of own experience, we assume that i would need to observe j obtaining a sequence of consecutive $\pi_t^j = 1$ of length $l^i = 2$, at least r^i times. Note also that i does not pay attention to j's payoff all the time (we keep the same probabilistic attention process described above), so that we need, in case of $l^i = 2$, $\Gamma_t^i \pi_t^j = \Gamma_{t=1}^i \pi_{t-1}^j = 1$ at least r^i times to become aware of the existence of the inter-temporal dependencies and conditional strategies.

An example may help. Consider the following history of the payoffs received by agent j and of the rounds in which agent i's paying attention.

t	20	21	22	23	24	25	26	27	28	29	30	31	32	33
π_t^j	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Γ_t^i	0	0	1	1	0	0	1	1	0	1	0	1	1	0

here i observed j obtaining a payoff of 1 in two consecutive rounds three times: in rounds 22-23, 26-27, and 31-32. If $r^i = 2$, we assume that from round 28 on, i will become aware of the existence of inter-temporal dependencies and start updating based on j's payoff, when i pays attention to it, as well.

If i is already aware of conditional strategies and has just experienced the (choice, payoff) sequence $(1, 1), (1, 0)$, while observed j obtaining payoff of 1 in the same two rounds, then i will infer (assuming he already knows that $\Pi(a)^i_t = \{0,1\}$ for $a \in \{1,2,3\}$ and $\Pi(4)^i_t = \{0.3\}$) that choosing either 2 or 3 instead of 1 after $(1,1)$ could have generated a payoff of 1. In fact, this type of inference is possible in our set-up only after i obtains a payoff of 1 in the previous round and 0 in the current one.[18] Notice also that this means that i will not use the information about j's payoff when s/he is not using a conditional strategy. We therefore assume that $A^i_t(a)$ evolves as follows:

$$A^i_{t+1}(a) = \begin{cases} \alpha^i A^i_t(a) + (1-\alpha^i)\pi^i_t & \text{if } a = a^i_t \\ A^i_t(a) & \text{otherwise} \end{cases}$$

On the other hand, if i is already using a conditional strategy (or has become aware of such strategies due to the observation of j's payoff) then,

$$B^i_{t+1}(a|h) = \begin{cases} \beta^i B^i_t(a|h) + (1-\beta^i)\pi^i_t & \text{if } a|h = a^i_t|h^i_{t-1} \\ \beta^i B^i_t(a|h) + (1-\beta^i)1 & \text{if } h = h^i_{t-1}, a \neq a^i_t, \pi^i_t = 0, \pi^i_{t-1} = \Gamma^i_{t-1}\pi^j_{t-1} = \Gamma^i_t \pi^j_t = 1, 1 \in \Pi^i_t(a) \\ B^i_t(a|h) & \text{otherwise} \end{cases}$$

Again, we are assuming that j's payoff information is used only if i has observed that j has obtained a payoff of 1 in two consecutive rounds. The choice of a conditional strategy in round t will be based on $B^i_t(a|h)$ just as in the baseline case.

5.2.3 The 'Choice Information' (CI) case

Here, intuitively, the decision will involve mimicking. We will, however, implement it within our modeling framework through the evolution of attractions to keep the modeling framework consistent across our four conditions instead of modeling it directly as agents copying the observed choice patterns. The basic idea will be that when agent i detects a pattern in the choices made by j, he will assume that j is doing so because it generates a high payoff. Thus, the attraction for conditional strategies will be updated with "presumed" high payoffs based on i's observation about j's choices (of course, given i's knowledge about the set of possible payoffs for each options and conditional

[18] After receiving a 0 payoff in the previous round, all choices (but 4) can result in both payoff of 0 or 1.

choices). This process operates in addition to the learning process described in our baseline model.

First i has to recognize patterns in j's choices. He does so when he observes that j has been making the same sequence of l^i consecutive choices, among three options other than 4, r^i times in the row.[19] Say i has observed j making the sequence of choices $3 \rightarrow 2$ twice in a row "recently." Here "recently" is defined as being among those recent block of two consecutive rounds in which i has paid attention to what j has chosen and j has chosen 3 in one round. If i has already become aware of inter-temporal dependencies, this will translate into i assuming that strategy $2|(3,1)$ will result in payoff of 1.

Again, an example may help. Consider the following history of choices made by agent j and of the rounds in which i was paying attention to it.

t	20	21	22	23	24	25	26	27	28	29	30	31	32	33
a_t^j	2	1	3	2	1	3	2	1	3	2	1	3	2	1
Γ_t^i	0	0	1	1	0	0	1	1	0	1	0	1	1	0

In this example, j is assumed to be following the best choice sequence in the setup, and i has paid attention to j's choice in rounds 22, 23, 26, 27, 29, 31, and 32. Of the three blocks of two consecutive rounds in which i paid attention, i.e., 22-23, 26-27, and 31-32, both the first and the last ones contain the choice sequence of $(3,2)$ by j. This will induce i to start making inferences about the conditional strategy $2|(3,1)$.

In the case where i has not yet become aware of conditional strategies, then the first time i notices the pattern in j's choice sequences, we assume that i then becomes aware of the existence of such strategies and starts learning about their performance.

Recall, however, that i does not observe j's payoff directly. Thus, we assume that when i is making unconditional choices, the observation of j's choice will not be used in the evolution of

[19] Excluding the sequence of choices involving 4, such as persistently choosing 4, is specific to our current experimental set up. But we do so here because we assume that agents will quickly learn that choosing 4 results in the payoff of 0.3.

$A_t^i(a)$; and that this will happen only when i's choices are conditional.

$$B_{t+1}^i(a|h) = \begin{cases} \beta^i B_t^i(a|h) + (1-\beta^i)\pi_t^i & \text{if } a|h = a_t^i|h_{t-1}^i \\ \beta^i B_t^i(a|h) + (1-\beta^i)1 & \text{if } a|h = (a_t^j|(a_{t-1}^j, 1)), \Gamma_t^i = \Gamma_{t-1}^i = 1, a_t^j \in \{1,2,3\} \\ B_t^i(a|h) & \text{otherwise} \end{cases}$$

The choice of a conditional strategy in round t will be based on $B_t^i(a|h)$ just as in the baseline case.

5.3 Simulation results

We simulate the above models using the set of parameter values $\left\{(\hat{\alpha}^i, \hat{\lambda}^i, \hat{r}^i)\right\}$ that best fit the baseline model, and assuming different values for $q^i = q$ for all i. That is, for each value of q and for each individual in each of 500 simulated pairs of agents, we randomly draw (with replacement) each of the parameter values from the set of parameters pertaining to the NI treatment. We have chosen to evaluate the goodness-of-fit of our models on the grounds of such out-of-sample predictions to overcome the identification problems that are inherent to the estimation of eight parameters for each pair of participants (i.e., $(\hat{\alpha}^i, \hat{\lambda}^i, \hat{r}^i)$ and q^i for each participant).

Figure 7 displays the simulated distributions of discovery times and total payoffs assuming that $q = 1$ (i.e., agents are assumed to always pay attention to the other's payoff and/or choice) along with the observed ones. The plots suggest that overall, participants discovered the hidden cycle somewhat faster and earned higher total payoffs than our simulated agents. However, running permutation tests indicates that differences between observed and simulated data samples are significant at $\alpha = 5\%$ only for total payoffs in the CPI treatment ($p = .024$, and $p = .058$ for discovery times). Otherwise, as the average statistics in Figure 7 indicate, our models provide a remarkably good out-of-sample fit to the data ($p = .500$ in PI and $p = .189$ in CI for discovery times, and $p = .482$ in PI and $p = .277$ in CI for total payoffs, two-tailed tests).

We proceed with cross-treatment comparisons of the models' simulations and the observations. The upper panel of Figure 8 displays the simulated distributions of discovery times (left panel)

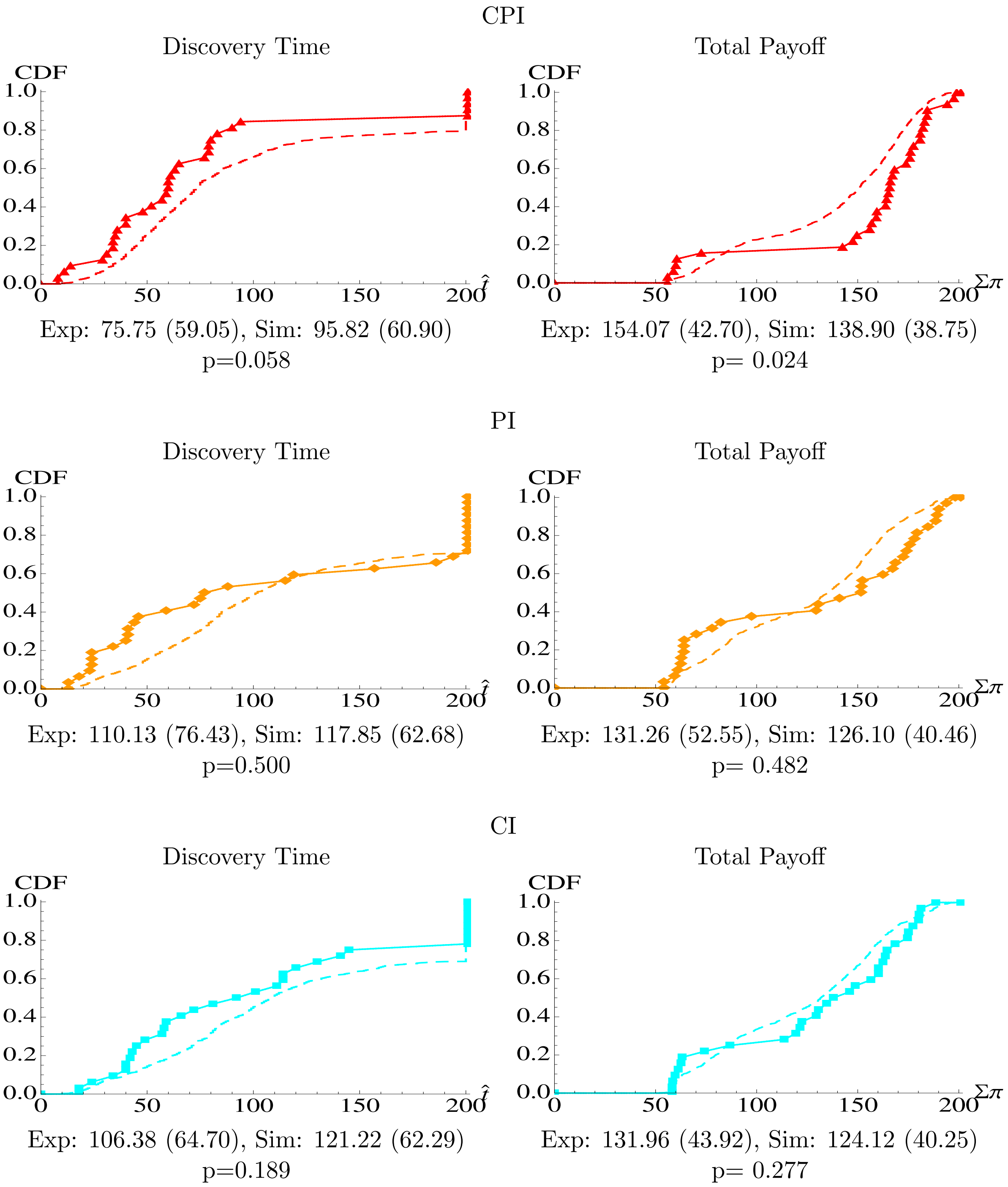
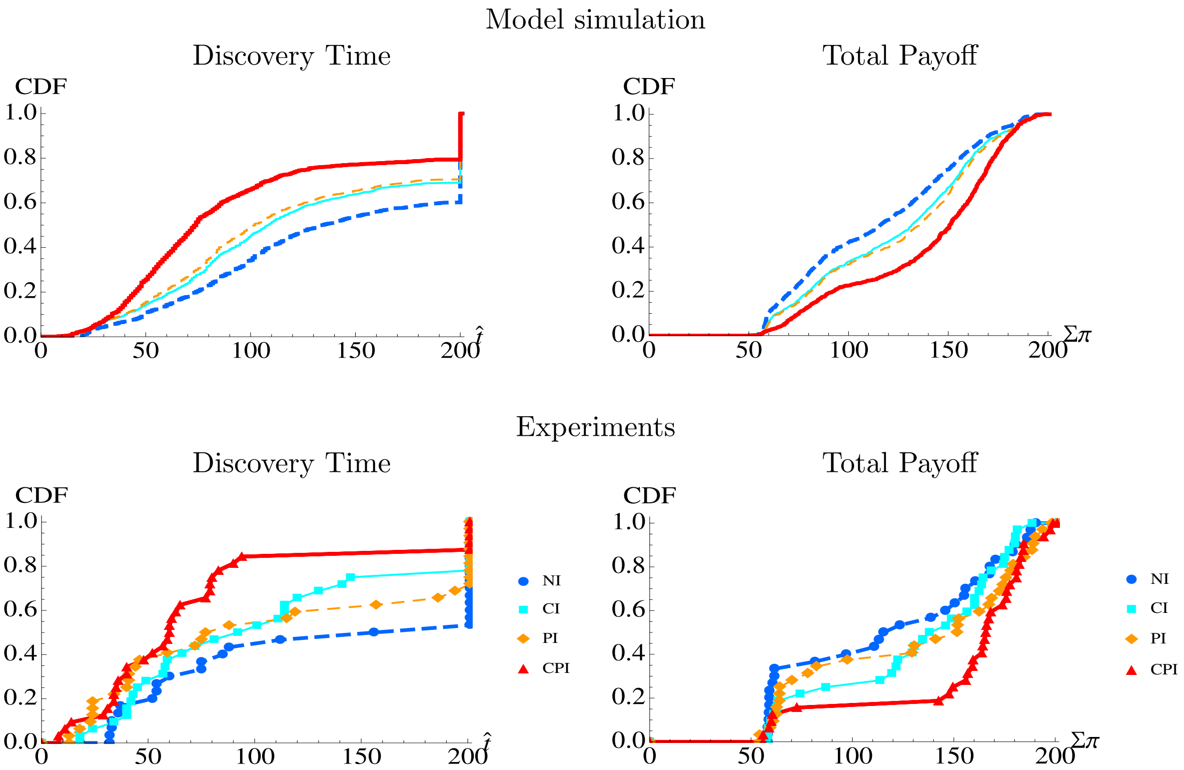


Figure 7: CDFs of discovery times and total payoffs.
Legend: Observed (solid), Simulations under $q^i = 1.0$ for all i (dashed). Mean (standard deviation) are also shown. P-values are based on permutation tests (for independent samples), two-tailed.



$p = 0.033$ (KW)

p-values	CI	PI	CPI
NI	0.065	0.114	0.001
CI		0.585	0.027
PI			0.025

$p = 0.025$ (KW)

p-values	CI	PI	CPI
NI	0.141	0.176	0.003
CI		0.522	0.023
PI			0.031

Figure 8: CDFs of discovery times and total payoffs (with p-values of Kruskal-Wallis tests for multiple sample comparisons and permutation tests for pair-wise comparisons (one-tailed)). Legend: NI (Thick dashed blue), CPI (Thick solid red), PI (Thin dashed orange), CI (Thin solid light blue).

and total payoffs (right panel), assuming $q = 1$, whereas the lower panel displays the observed distributions along with the outcomes of Kruskal-Wallis (KW) and permutation tests that assess their equivalence. There are two things worth noting here. First, the simulated distributions clearly confirm the conjecture that the more items of information one has about the other participant, the faster the hidden pattern will be discovered and the higher the total payoffs will be. This is confirmed by the observed distributions and test statistics for the NI and the CPI treatments but not by those of the partial information treatments CI and PI which are not significantly different from the NI treatment (p-values > 0.05). Second, the simulated distributions for the CI and PI treatments are virtually identical and the observed ones are not significantly different which suggests that the type of information received about the other participant does not significantly affect the variables of interest. In the light of Sonsino (1997)'s claim that the repeated observation of the same pattern of choices over a long period of time must be recognized by the observer (cf. Footnote 3), these findings raise the question as to whether observing the others' choices is cognitively equivalent to observing the other's payoff outcomes. On the one hand, the lack of significant difference between the simulated distributions of the CI and PI treatments and the empirical support of this prediction suggest that this is indeed the case. However, as we noted when presenting the pattern-learning model for the PI treatment, one could argue that observing another's payoff pattern is not as informative as observing another's choice pattern since the latter immediately calls for an imitation of such behaviour that might lead to the discovery of the hidden pattern. This is offset by the fact that observing a series of payoffs of 1, for example, reveals that there is a pattern of actions which leads to these high payoffs and thus stimulates the individual to search for such a pattern. This, in turn, would suggest that, if there is an informational difference, our observational pattern-learning model fails to capture it. On the other hand, the lack of significant difference between the observed distributions of CI and NI and of PI and NI suggest that the provision of information on past choices or past payoffs is in fact not sufficient to facilitate pattern-discovery in our framework. Here we recall that our descriptive data analyses indicated a considerable heterogeneity in subjects' behaviour which, in turn, may have interfered with our predictions.

5.4 Does observation facilitate discovery?

Although it is quite intuitive that observing the choices and/or payoffs of another person who has discovered the hidden pattern should help an observer who has not yet found the pattern, it is less clear whether observing the choices and/or payoffs of another participant who has not found any pattern is equally helpful. We investigate this by first identifying in each pair of simulated agents, the one who discovered the hidden pattern first, henceforth the Early finder, or second, henceforth the Late finder. If the hidden pattern was not discovered, then the labels are randomly assigned. For the baseline treatment, we randomly match agents in pairs and assign labels by comparing their respective discovery times so as to get some benchmark for Early and Late finders who do not observe any other agent's outcome(s).

Figure 9 shows the distributions of discovery times of each type of simulated agents for four values of $q \in \{0.4, 0.6, 0.8, 1\}$.

The simulations suggest that if $q = 1$ (as we assumed above), the distributions of discovery times for both Early and Late finders will be ordered as CPI $<$ CI $\approx$ PI $<$ NI. However, for lower values of q such as $q = 0.6$, the distributions of discovery time for the Early finders in the information treatments (CI, PI, CPI) become similar, while those of Late finders maintain the same ordering as when $q = 1$. And if $q = 0.4$, the discovery times of Early finders are basically the same across treatments whereas for Late finders, only CPI results in a markedly earlier discovery compared to the other treatments.

We proceed by displaying the observed distributions of discovery times of Early and Late finders in Figure 10. The plots and KW test statistics for Late finders indicate that they greatly benefited from the additional information; just as in our simulations with $q \geq 0.6$. Those in CPI benefited most and significantly more than those in CI or PI whose distributions are not significantly different. When compared to NI, the plots and test statistics also indicate that the partial information feedback of these treatments significantly improved their discovery times. Interestingly, the observed distributions of Late finders in CI also suggest that when compared to PI, discovery times sharply decrease from round 110 onwards which is in line with the conjecture that the CI treatment should make it easier to discover the pattern once the other has discovered it and exhibits a cyclical

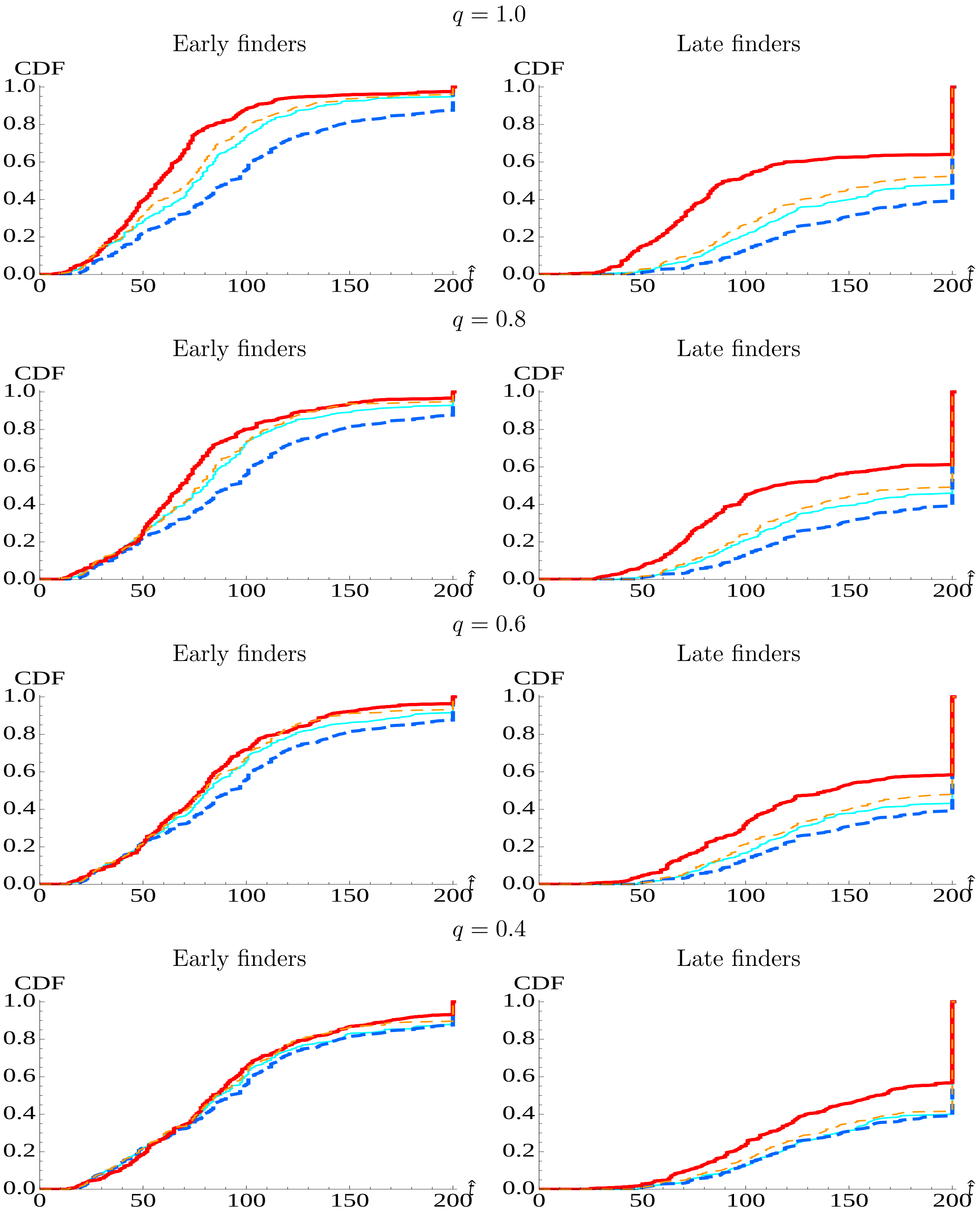
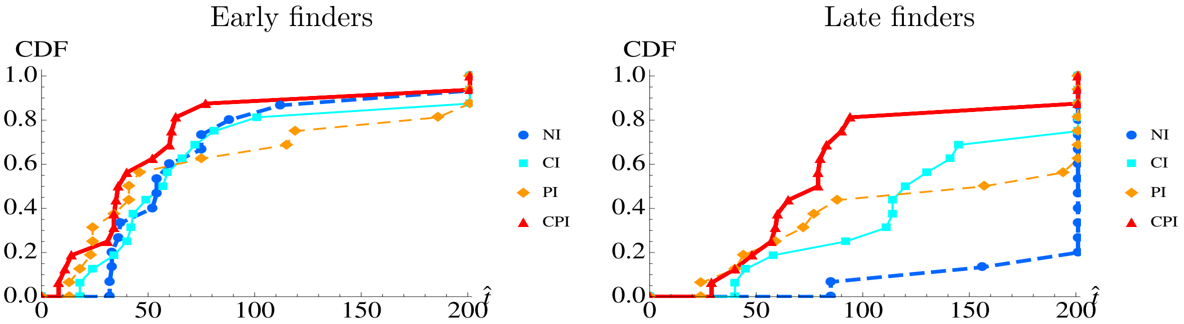


Figure 9: CDFs of the simulated discovery times of Early and Late finders for three values of q for all the treatments. $q = 1.0$ (1st row), $q = 0.8$ (2nd row), $q = 0.6$ (3rd row), and $q = 0.4$ (4th row). Legend: NI (Thick dashed blue), CPI (Thick solid red), PI (Thin dashed orange), CI (Thin solid light blue).



$p = 0.527$ (KW)

p-values	CI	PI	CPI
NI	0.575	0.647	0.221
CI		0.575	0.172
PI			0.147

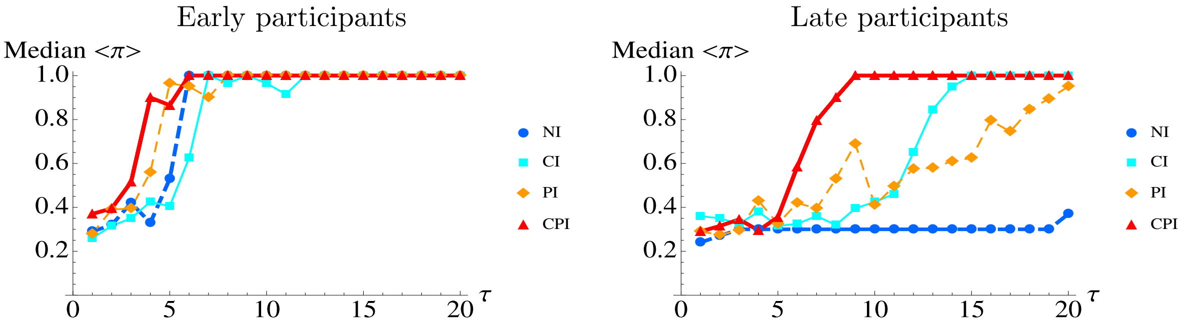
$p < 0.001$ (KW)

p-values	CI	PI	CPI
NI	0.001	0.005	< 0.001
CI		0.551	0.027
PI			0.035

Figure 10: CDFs of discovery times of Early and Late finders in the experiments (with p-values of Kruskal-Wallis tests for multiple sample comparisons and permutation tests for pair-wise comparisons (one-tailed)).
Legend: NI (Thick dashed blue), CPI (Thick solid red), PI (Thin dashed orange), CI (Thin solid light blue).

choice pattern. As for the Early finders, we find no significant difference across treatments, as in our model simulation for $q \leq 0.6$, so that the additional information provided had no significant effect on their performance.

To further determine who benefited most from the information provided, we track the evolution of average payoffs of Early and Late finders in each treatment. Figure 11 reports their respective median of round average payoffs for each batch of 10 rounds in each treatment. The plots and KW test statistics suggest that round average payoffs are not significantly different across treatments for the first five batches and are actually very similar across types (i.e., Early and Late). Differences emerge from the sixth batch onwards: the round average payoffs of Late finders vary widely across treatments whereas those of Early finders become identical. This confirms that the information provided did not affect Early finders whereas it significantly helped Late finders. Note also that the median average round payoffs in CI sharply increase between the eleventh and fifteenth batch, which is again in line with the claim that discovering a pattern is easier as a Late finder in CI than in PI because one can simply copy the systematic behavior of the Early finder.



p-values from Kruskal-Wallis test

τ	1	2	3	4	5	6	7	8	9	10
Early participants	0.249	0.709	0.814	0.416	0.801	0.288	0.430	0.261	0.852	0.300
Late participants	0.098	0.204	0.592	0.067	0.981	**0.008**	**0.044**	**0.039**	**0.004**	**0.018**
τ	11	12	13	14	15	16	17	18	19	20
Early participants	0.126	0.317	0.638	0.852	0.693	0.759	0.690	0.511	0.885	0.676
Late participants	**0.003**	**0.019**	**0.014**	**0.005**	**0.014**	**0.012**	**0.008**	**0.032**	**0.010**	**0.054**

Figure 11: Dynamics of median round average payoff (over block of 10 rounds) of Early and Late participants (with p-values of Kruskal-Wallis tests).

To this extent, the models' general prediction (with $q < 1$) that the more information about the other participant's last decision and/or reward accelerates discovery appears to hold mainly for Late finders, and the conjecture that it should be even faster in CI than in PI seems to hold only once Late finders accumulated some experience of the game they were playing. So we have a partial answer to the relative speed of discovery in CI as opposed to PI which we discussed earlier but which is not directly incorporated in our model.

6 Conclusion

We report on a series of bandit experiments that manipulate the end-of-round information feedback to assess the effect of observational learning on the discovery of the bandits' hidden payoff structure. We considered four treatments, each involving four options and consisting of 200 rounds of play. The payoffs of the first three options have an undisclosed deterministic temporal structure which generates a payoff of 1 in each round, provided that the options are chosen in the right sequence, whereas the fourth generates a sure payoff of 0.3.

In a baseline treatment (NI), participants played in isolation, i.e., with no information on other's play or outcomes. While some participants were surprisingly fast in "discovering" the hidden mechanism in this treatment, half of them failed to do so by the end of the experiment. Most participants initially explored among four different options. At some point, however, those who were successful seem to have had an "aha" moment, and started to reinforce on sequences of actions taken in the past. We propose a simple reinforcement model that captures such pattern-learning and find that it reproduces the distributional properties of the observed discovery times and total payoffs remarkably well.

In the three other treatments, participants were randomly matched in fixed pairs and received information either on the other's last choice (CI), the other's last payoff (PI) or on both the other's last choice and payoff (CPI). These treatments allow us to assess the effect of observational learning on the participants' discovery of the bandits' payoff pattern and the data confirms earlier findings pertaining to game theoretic settings that the provision of information on both the other's last choice and payoff mostly accelerates discovery whereas the provision of partial information (i.e., the other's last choice or last payoff) does not when compared to the baseline treatment (NI). We extend our basic model to capture the behaviour observed in these treatments and find that the discovery times are well supported by out-of-sample predictions in the partial information treatments but not in the CPI one where participants performed even better than predicted. Such a gap between predictions and observations in this treatment may be evidence for a compounding effect of information that is not captured by our model. On the other hand, the irrelevance of the type of information (choice or payoff) on the participants' discovery times in CI and PI may result from their heterogeneous traits. We therefore identify Early and Late finders within each pair and find that the information provided significantly improves the discovery times of Late finders in all three treatments while leaving those of Early finders largely unaffected. This is supported by the model's out-of-sample predictions and/or by an *ex post* adjustment of the probability of paying attention to the other's outcome(s).

To the best of our knowledge, our approach is the first to study this type of pattern-learning with and without observational learning in an interaction-free environment. We believe that our predictions could be further refined and stress-tested with further experiments that would control the size (or presence) of the safe bandit-option, the hidden payoff pattern and/or the length of the game's history provided in the information treatments.

Finally, our investigation relates to recent research in the neurosciences which investigates which are the mental processes at work that lead individuals to learn how to understand and operate in their environment. There, it is argued that different mechanisms operate in the brain when an individual switches from exploiting the information about the various options he has tried to exploring new ones (see Cohen et al., 2007; Laureiro-Martínez et al., 2014, 2015).[20] One approach has been to identify the neural processes at work at "aha" or, referring to Archimedes, the "eureka" moment when an individual suddenly perceives the answer to a problem s/he has been wrestling with. There is now a substantial literature trying to identify the changes in neural activity that take place when such a realisation occurs (see Auble et al., 1979; Kounios et al., 2008; Topolinski and Reber, 2010). However, this literature has focused on solving a well-defined problem to which it is known that there is a "solution" such as the Rubik's cube or a mathematical puzzle. To better understand the mechanisms involved in producing an "aha" moment, when subjects are facing a situation without known solution would be of considerable interest.

References

Agarwal, A., D. Hsu, S. Kale, J. Langford, L. Li, and R. E. Schapire (2014): "Taming the monster: A fast and simple algorithm for contextual bandits," in *In Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 1638–1646.

Armantier, O. (2004): "Does observation influence learning?" *Games and Economic Behavior*, 46, 221–239.

[20] Not only are different networks of neurons activated in the two cases but the process of switching is different depending on the activity that an individual normally pursues. The authors compared two groups of subjects, managers and entrepreneurs, and found that the threshold at which a manager passed to exploration is significantly higher than that of an entrepreneur. In other words, entrepreneurs tend to function in the exploratory mode more often than managers. Indeed, the distinction that emerges between people with different professions suggests that this process is also present in longer term situations.

Arrow, K. J. and J. R. Green (1973): "Notes on Expectations Equilibria in Bayesian Settings," Working Paper 33, Institute for Mathematical Studies in the Social Sciences, Stanford University.

Auble, P. M., J. J. Franks, and J. Salvatore A. Soraci (1979): "Effort toward comprehension: Elaboration or "aha!"?" *Memory & Cognition*, 7.

Auer, P. (2002): "Using Confidence Bounds for Exploitation-Exploration Trade-offs," *Journal of Machine Learning Research*, 3, 397–422.

Bandura, A. and F. J. McDonald (1963): "Influence of social reinforcement and the behavior of models in shaping children's judgment," *The Journal of Abnormal and Social Psychology*, 67, 274–281.

Bandura, A., D. Ross, and S. A. Ross (1963): "Vicarious reinforcement and imitative learning," *The Journal of Abnormal and Social Psychology*, 67, 601–607.

Banks, J., M. Olson, and D. Porter (1997): "An experimental analysis of the bandit problem," *Economic Theory*, 10, 55–77.

Bayer, R.-C. and H. Wu (2016): "Do we learn from our own experience or from observing others?" mimeo, National University of Singapore.

Beygelzimer, A., J. Langford, L. Li, L. Reyzin, and R. E. Schapire (2011): "Contextual bandit algorithms with supervised learning guarantees," in *In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AIS-TATS).*

Bossan, B., O. Jann, and P. Hammerstein (2015): "The evolution of social learning and its economic consequences," *Journal of Economic Behavior and Organization*, 112, 266–288.

Bouchaud, J.-P. (2013): "Crises and Collective Socio-Economic Phenomena: Simple Models and Challenges," *Journal of Statistical Physics*, 151, 567–606.

Bray, M. (1982): "Learning, estimation and stability of rational expectations," *Journal of Economic Theory*, 26, 318–339.

Brown, S., M. Steyvers, and E.-J. Wagenmakers (2009): "Observing evidence accumulation during multi-alternative decisions," *Journal of Mathematical Psychology*, 53, 453–462.

Burke, C. J., P. N. Tobler, M. Baddeley, and W. Schultz (2010): "Neural mechanisms of observation learning," *Proceedings of the National Academy of Science, U.S.A.*, 107, 1443114436.

Bush, R. R. and F. Mosteller (1951): "A mathematical model for simple learning," *The Pyschological Review*, 58, 313–323.

Camerer, C. and T.-H. Ho (1999): "Experience-weighted attraction learning in normal form games," *Econometrica*, 67, 827–874.

Cohen, J. D., S. M. McClure, and A. J. Yu (2007): "Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration," *Philosophical Transactions of the Royal Society B*, 362, 933–942.

Efferson, C., P. J. Richerson, R. McElreath, M. Lubell, E. Edsten, T. M. Waring, B. Paciotti, and W. Baum (2007): "Learning, productivity, and noise: an experimental study of cultural transmission on the Bolivian Altiplano," *Evolution and Human Behavior*, 28, 11–17.

Erev, I. and A. E. Roth (1998): "Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria," *American Economic Review*, 88, 848–881.

Fischbacher, U. (2007): "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10, 171–178.

Fryer, R. and P. Harms (2017): "Two-Armed Restless Bandits with Imperfect Information: Stochastic Control and Indexability," rxiv:1506.07291v2, arXiv.org.

Fryling, M. J., C. Johnston, and L. J. Hayes (2011): "Understanding Observational Learning: An Interbehavioral Approach," *The Analysis of Verbal Behavior*, 27, 191–203.

Garivier, A. and E. Kaufmann (2016): "Optimal Best Arm Identification with Fixed Confidence," Tech. Rep. arXiv:1602.04589v2, arXiv.org.

GREINER, B. (2015): "Subject pool recruitment procedures: organizing experiments with ORSEE," *Journal of the Economic Science Association*, 1, 114–125.

HILLS, T. T., P. M. TODD, D. LAZER, A. D. REDISH, I. D. COUZIN, AND "THE COGNITIVE SEARCH RESEARCH GROUP" (2015): "Exploration versus exploitation in space, mind, and society," *Trends in Cognitive Sciences*, 19, 46–54.

HOLT, C. A. AND S. K. LAURY (2002): "Risk Aversion and Incentive Effects," *American Economic Review*, 92, 1644–1655.

HU, Y., Y. KAYABA, AND M. SHUM (2013): "Nonparametric learning rules from bandit experiments: The eyes have it!" *Games and Economic Behavior*, 81, 215–231.

KIRMAN, A. (1975): "Learning by firms about demand conditions," in *Adaptive Economic Models*, ed. by R. H. Day and T. Groves, Academic Press, 137–156.

——— (1983): "Mistaken beliefs and resultant equilibria," in *Individual Forecasting and Collective Outcomes "Rational expectations" examined*, ed. by R. Frydman and E. S. Phelps, Cambridge, UK: Cambridge University Press, chap. 8, 147–168.

KOUNIOS, J., J. I. FLECK, D. L. GREEN, L. PAYNE, J. L. STEVENSON, E. M. BOWDEN, AND M. JUNG-BEEMAN (2008): "The Origins of Insight in Resting-State Brain Activity," *Neuropsychologia*, 46, 281–291.

LAIRD, P. AND R. SAUL (1994): "Discrete Sequence Prediction and Its Applications," *Machine Learning*, 15, 43–68.

LAUREIRO-MARTÍNEZ, D., S. BRUSONI, AND N. C. ANDMAURIZIO ZOLLO (2015): "Understanding the exploration–exploitation dilemma: An fMRI study of attention control and decision-making performance," *Strategic Management Journal*, 36, 319–338.

LAUREIRO-MARTÍNEZ, D., N. CANESSA, S. BRUSONI, M. ZOLLO, T. HARE, F. ALEMANNO, AND S. F. CAPPA (2014): "Frontopolar cortex and decision-making efficiency: comparing brain activity of experts with different professional background during an exploration-exploitation task," *Frontiers in Human Neuroscience*, 7, 1–10.

Marchiori, D. and M. Warglien (2008): "Predicting human interactive learning by regret-driven neural networks," *Science*, 319, 1111–1113.

McElreath, R., M. Lubell, P. J. Richerson, T. M. Waring, W. Baum, E. Edsten, C. Efferson, and B. Paciotti (2005): "Applying evolutionary models to the laboratory study of social learning," *Evolution and Human Behavior*, 26.

McMahan, H. B. and M. J. Streeter (2009): "Tighter Bounds for Multi-Armed Bandits with Expert Advice," in *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*.

Nadal, J.-P., O. Chenevez, G. Weisbuch, and A. Kirman (1998): "A Formal Approach to Market Organization: Choice Functions, Mean Field Approximation and Maximum Entropy Principle," in *Advances in Self-Organization and Evolutionary Economics*, ed. by J. Lesourne and A. Orléan, London: Economica, 149 – 159.

Nedic, A., D. Tomlin, P. Holmes, D. A. Prentice, and J. D. Cohen (2012): "A Decision Task in a Social Context: Human Experiments, Models, and Analyses of Behavioral Data," in *Proceedings of the IEEE. Interaction Dynamics: The Interface of Humans and Smart Machines*, ed. by J. Baillieul, N. E. Leonard, and K. A. Morgansen, 713 – 733.

Smith, L. and P. N. Sørensen (2011): "Observational learning," in *The New Palgrave Dictionary of Economics*, ed. by S. N. Durlauf and L. E. Blume, Palgrave Macmillan, online edition ed.

Sonsino, D. (1997): "Learning to learng, pattern recognition, and Nash equilibrium," *Games and Economic Behavior*, 18, 286–331.

Spiliopoulos, L. (2012): "Pattern recognition and subjective belief learning in a repeated constant-sum game," *Games and Economic Behavior*, 75, 921–935.

——— (2013): "Beyond fictitious play beliefs: Incorporating pattern recognition and similarity matching," *Games and Economic Behavior*, 81, 69–85.

Steyvers, M., M. D. Lee, and E.-J. Wagenmakers (2009): "A Bayesian Analysis of Human Decision-Making on Bandit Problem," *Journal of Mathematical Psychology*, 53, 168–179.

Topolinski, S. and R. Reber (2010): "Gaining Insight Into the "Aha" Experience," *Current Directions in Psychological Science*, 19, 402–405.

Woodford, M. (1990): "Learning to believe in sunspots," *Econometrica*, 58, 277–307.

A Instructions for treatments CPI, PI and CI

Welcome to the ASBLab.

If you read the following instructions carefully, you can, depending on your decisions, earn a considerable amount of money. It is therefore very important that you read these instructions carefully. The instructions are the same for all participants.

It is prohibited to communicate with the other participants during the experiment. If you have a question at any time raise your hand and the experimenter will come to your desk to answer it. Please switch off your mobile phone or any other devices which may disturb the experiment. Please use the computer only for entering your decisions. Please do not start or end any programs, and do not change any settings.

This Experiment

You are about to participate in an experiment which consists of two parts.

The first part consists of 200 rounds of play. The second part will be explained to you when you finished the first part.

First Part

Task

In each of the 200 rounds of play, you are asked to choose one of 4 one-armed bandits. Once you made your choice, you will be informed and will receive the bandits' payoff (in Experimental Currency Units, ECUs). The experiment then proceeds to the next round.

Information

You are not aware of the payoffs that you may receive from each of these bandits but you are told that a bandits' payoff outcome in one round does not depend on which bandit you chose in the previous round.

You are not allowed to collect any written information on the bandits' payoff outcomes.

== [CPI treatment only] ==

Throughout the experiment, you will be matched with one other participant. At the end of each round, you will get to know this participant's bandit choice and the payoff outcome of that choice, and s/he will get to know your bandit choice and the associated payoff. Note that the information displayed is about bandit choices and the payoff outcomes associated to these choices.

== [CI treatment only] ==

Throughout the experiment, you will be matched with one other participant. At the end of each round, you will get to know this participant's bandit choice, and s/he will get to know your bandit choice.

== [PI treatment only] ==

Throughout the experiment, you will be matched with one other participant. At the end of each round, you will get to know the payoff outcome for this participant's, and s/he will get to know your payoff outcome.

Payment

Your reward from participating to this first part will be equal the sum of the payoffs that you realised, and this sum will be converted to the rate of $0.2 per ECU and individually paid to you.

Second Part

Task

In the second part of the experiment, you are asked 10 times to choose between "Option A" and "Option B." Please read carefully the questions asked. (The questions were displayed in sequence and were phrased as in Table 2 below)

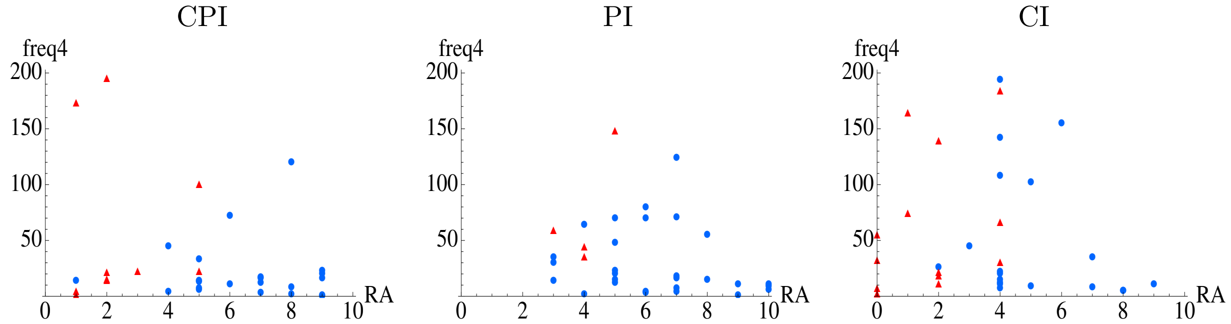
Payment

Once you have answered all questions, your reward from participating to this second part will be determined by 1) your answer to one of these ten questions and 2) by chance. The computer will randomly select one of the ten questions that you have answered and you will be rewarded according to your decision (ie., Option A or Option B) for that question.

Even though you will make ten decisions, only one of these will end up affecting your earnings, but you will not know in advance which decision will be used. Obviously, each decision has an equal chance of being used in the end.

Table 2: List of questions fro Holt and Loury task.

	Option A	**Option B**
1.	A$2.0 with 10% chance or A$1.60 with 90% chance	A$3.85 with 10% chance or A$0.1 with 90% chance
2.	A$2.0 with 20% chance or A$1.60 with 80% chance	A$3.85 with 20% chance or A$0.1 with 80% chance
3.	A$2.0 with 30% chance or A$1.60 with 70% chance	A$3.85 with 30% chance or A$0.1 with 70% chance
4.	A$2.0 with 40% chance or A$1.60 with 60% chance	A$3.85 with 40% chance or A$0.1 with 60% chance
5.	A$2.0 with 50% chance or A$1.60 with 50% chance	A$3.85 with 50% chance or A$0.1 with 50% chance
6.	A$2.0 with 60% chance or A$1.60 with 40% chance	A$3.85 with 60% chance or A$0.1 with 40% chance
7.	A$2.0 with 70% chance or A$1.60 with 30% chance	A$3.85 with 70% chance or A$0.1 with 30% chance
8.	A$2.0 with 80% chance or A$1.60 with 20% chance	A$3.85 with 80% chance or A$0.1 with 20% chance
9.	A$2.0 with 90% chance or A$1.60 with 10% chance	A$3.85 with 90% chance or A$0.1 with 10% chance
10.	A$2.0 with 100% chance or A$1.60 with 0% chance	A$3.85 with 100% chance or A$0.1 with 0% chance



Without triangles

$N = 22$	$N = 28$	$N = 19$
$\rho = -0.094$ ($p = 0.680$)	$\rho = -0.282$ ($p = 0.148$)	$\rho = -0.413$ ($p = 0.079$)

With triangles ($N = 32$ each in three treatments)

$\rho = -0.151$ ($p = 0.414$)	$\rho = -0.331$ ($p = 0.064$)	$\rho = -0.145$ ($p = 0.431$)

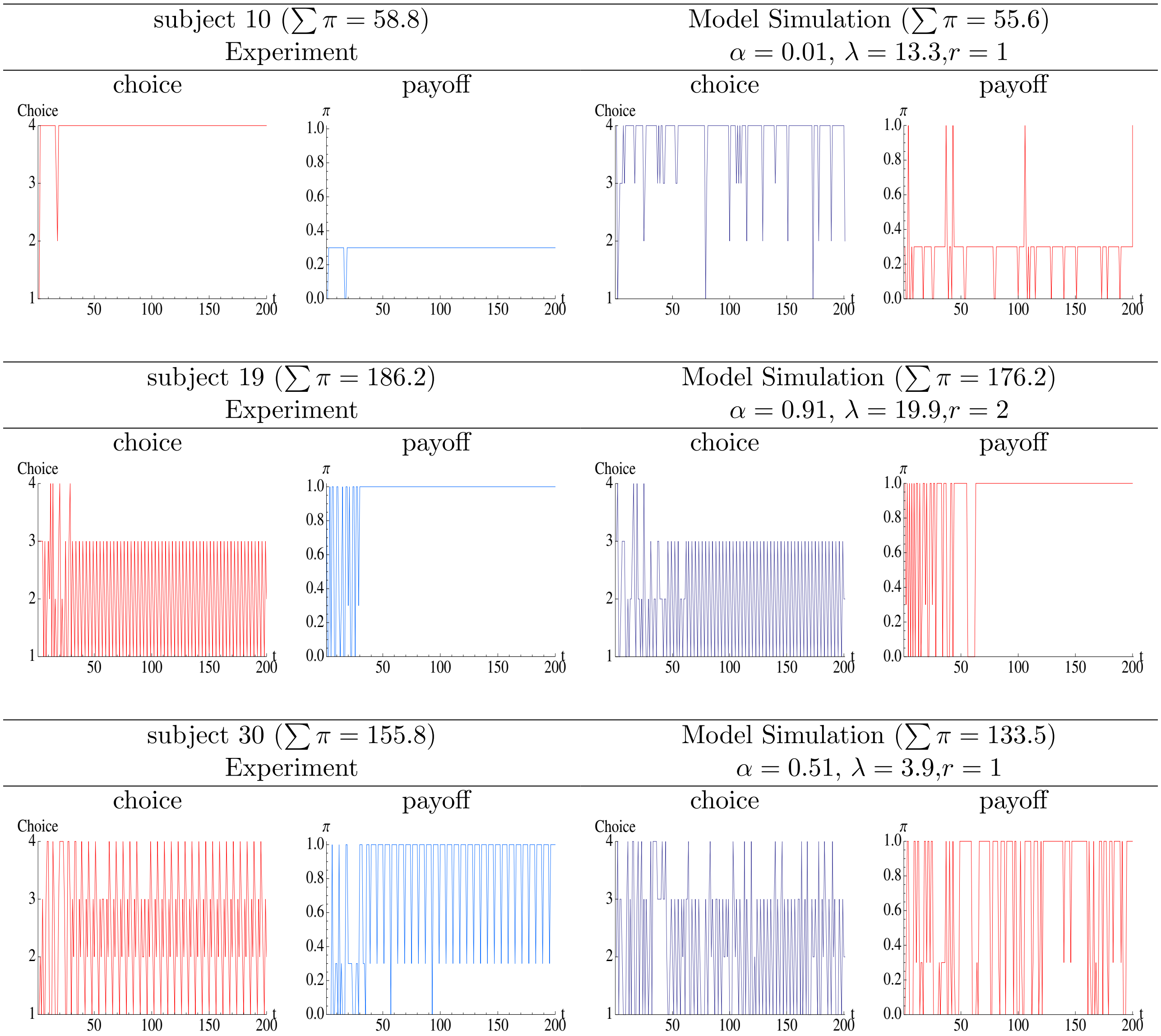
Figure 12: Risk Aversion (x-axis) *vs* Frequency of Safe Option (y-axis). Triangle: subjects who have switched multiple times in Holt and Laury (2002) questionnaire.

B Risk aversion and choices in paired treatments.

In this appendix we report on the correlation between the observed frequencies of safe choices and the participants' risk aversion in treatments CPI, CI and PI. The scatter plots in Figure 12 suggest no clear tendency in this regard, and since the reported correlation coefficients are not statistically significant (p-values $> .05$) we reject the conjecture that behaviour in these information treatments is driven by risk aversion just as in the case of NI treatment.

C Examples of fitted model simulation

Figure 13 shows the time series of observed and simulated choices (left panel) and payoffs (right panel) for three of the six participants displayed in Figure 1. Our model replicates quite well the behaviour of Subject 19 who discovered the hidden cycle within 50 rounds, and of Subject 10 who settled for the safe choice from early on. This, however, is less so for Subject 30 who ended-up using an almost optimal cyclical pattern of length 6 (i.e., 3, 2, 1, 3, 2, 4).


subject 10 ($\sum \pi = 58.8$)
Experiment

Model Simulation ($\sum \pi = 55.6$)
$\alpha = 0.01,\ \lambda = 13.3, r = 1$

subject 19 ($\sum \pi = 186.2$)
Experiment

Model Simulation ($\sum \pi = 176.2$)
$\alpha = 0.91,\ \lambda = 19.9, r = 2$

subject 30 ($\sum \pi = 155.8$)
Experiment

Model Simulation ($\sum \pi = 133.5$)
$\alpha = 0.51,\ \lambda = 3.9, r = 1$

choice

payoff

Figure 13: Comparison of the experimental data and simulated model for three subjects.